

# MoA: Mixture of Aggregators Improves Slide-Level Diagnosis in Computational Pathology


**Fatih Ozluedik**<sup>\*1</sup>

FATIH.OEZLUEGEDIK@HELMHOLTZ-MUNICH.DE

<sup>1</sup> *Institute of AI for Health, Helmholtz Munich, Munich, Germany*

**Muhammed Furkan Dasdelen**<sup>\*1</sup> 

FURKAN.DASDELEN@HELMHOLTZ-MUNICH.DE

**Rao Muhammad Umer**<sup>1</sup> 

UMER.RAO@HELMHOLTZ-MUNICH.DE

**Carsten Marr**<sup>1,2</sup> 

CARSTEN.MARR@HELMHOLTZ-MUNICH.DE

<sup>2</sup> *Department of Medicine III, Ludwig-Maximilian-University Hospital, Munich, Germany*

**Editors:** Under Review for MIDL 2026

## Abstract

Multiple instance learning (MIL) is the standard for learning slide-level representations from whole slide images (WSIs), typically using a single attention-based aggregator to pool instance features. However, a single aggregator can struggle to capture the diverse morphological patterns in heterogeneous pathology and cytology data, where different diseases may demand different pooling behaviours. We propose a mixture-of-aggregators framework that models complementary aspects of instance distributions in histology and hematologic cytology. A router with top-2 gating dynamically selects the most relevant aggregators per slide, and their outputs are fused into a patient-level representation. To avoid collapse to a single dominant expert, we add a load-balancing loss and Gumbel noise on the router logits to promote the use of multiple aggregators. We extensively evaluate our method on 19 different tasks derived from 16 datasets including histology and hematologic cytology. Compared to single-aggregator baselines, our approach improves diagnostic prediction accuracy by an average of 4.5% over ABMIL and 12.6% over TransMIL across all tasks. Beyond performance, our analysis shows that different aggregators attend to distinct, disease-specific instance distributions, providing interpretable insights into the diagnostic process.

**Keywords:** computational pathology, multiple instance learning, cytology

## 1. Introduction

Pathology whole-slide images (WSIs) are indispensable for cancer diagnosis, but their manual assessment is time-consuming and highly dependent on expert experience. With the advent of high-throughput slide digitization, AI-based approaches have been introduced to support tumor detection, grading, morphological and molecular subtyping, and even survival prediction (Chen et al., 2024; Lu et al., 2023; Bilal et al., 2021; Vorontsov et al., 2024; Li et al., 2023). These tools can reduce the workload of pathologists while enabling faster and more standardized diagnostic outputs (Bulten et al., 2021; Dy et al., 2024; Steiner et al., 2018; Janowczyk et al., 2019). WSIs are extremely large, often reaching gigapixel resolution, yet typically come with only slide-level labels. To bridge this gap, multiple instance learning (MIL) has become the standard paradigm. In MIL, a WSI is partitioned into non-overlapping patches, each patch is encoded into a feature representation, and an aggregator combines these patch-level features into a slide-level embedding for downstream

---

\* Contributed equally

tasks (Li et al., 2021; Lu et al., 2021; Campanella et al., 2019; Ilse et al., 2018; Shao et al., 2021; Ding et al., 2024).

The success of MIL depends critically on both the quality of instance encoding and the design of the aggregator. Modern pathology foundation models provide strong instance-level features (Filiot et al., 2024; Biopimus, 2025; Zimmermann et al., 2024; Chen et al., 2024; Lu et al., 2024), but aggregation remains challenging (Chen et al., 2024; Ding et al., 2024). Most current MIL methods focus on identifying a small subset of diagnostically relevant patches, which is effective when the presence of a single pattern is sufficient for diagnosis. However, many diseases are defined not only by the presence of specific cell types or structures, but also by their distribution and relative frequency within the slide. Conventional single-aggregator approaches are prone to fail to capture these subtler distributional patterns, leading to a loss of diagnostically important information (Lu et al., 2021; Li et al., 2021; Shao et al., 2021).

Mixture-of-experts strategies, widely adopted in large language models, show that dividing responsibility across multiple specialized components allows the system to model diverse tasks and distributions more effectively (Jacobs et al., 1991; Shazeer et al., 2017; Riquelme et al., 2021). Inspired by this idea, we propose a mixture of aggregators for computational pathology. Within a single pipeline, multiple aggregators can learn complementary aspects of slide composition—some focusing on highly discriminative instances, others capturing broader distributional signals. We hypothesize that such diversity enables the model to represent distinct disease-specific distributions more faithfully, leading to improved diagnostic performance and better alignment with clinical reasoning. The main contributions of our work are: (i) Instead of a single aggregator, we train multiple aggregators in a MIL pipeline using a routing strategy that weights each aggregator’s contribution. (ii) Our pipeline supports diverse aggregator architectures and improves performance. (iii) We show that each aggregator captures distinct, diagnostically relevant, and complementary instance distributions.

## 2. Related work

### 2.1. Aggregators in multiple instance learning

Early multiple instance learning (MIL) approaches for whole-slide images (WSIs) employed non-parametric, permutation-invariant pooling functions—such as mean, max, and log-sum-exp (LSE)—to compress instance features into slide-level representations (Campanella et al., 2019; Ilse et al., 2020; Keshvarikhojasteh, 2025). While simple and efficient, these fixed functions have limited capacity to adapt to data.

A major advance beyond static pooling mechanisms was the introduction of Attention-based Multiple Instance Learning (ABMIL) (Ilse et al., 2018; Sadafi et al., 2020). ABMIL learns instance-specific attention weights and computes a weighted average of patch embeddings, enabling more flexible slide-level predictions and interpretable instance-level heatmaps. Building on ABMIL, several extensions have been proposed. Clustering-constrained Attention MIL (CLAM) incorporates instance-level clustering to promote diverse class-specific prototypes (Lu et al., 2021), while Dual-stream MIL (DSMIL) couples an instance-discriminative stream with a bag-level stream through contrastive alignment (Li et al., 2021). More recently, Transformer-based aggregators such as TransMIL apply self-attention

across patches, explicitly modeling inter-instance relationships and often achieving improved whole-slide accuracy (Shao et al., 2021).

Despite their architectural differences, these MIL approaches share a common bottleneck in how they form the final slide-level representation. In most implementations—including ABMIL, DSMIL, and Transformer-based MIL models—the bag is ultimately reduced to a pooling operation. This is typically realized either through attention pooling or through a classification token (CLS) whose final hidden state  $h_{\text{cls}}^{(L)}$  summarizes the entire set. Such readouts are mathematically equivalent to Pooling by Multi-Head Attention (PMA) with  $k = 1$  in Set Transformers (Lee et al., 2019), and fall under the Deep Sets formulation (Zaheer et al., 2017). In essence, the model relies on a single learned query vector that attends over all patches, producing a learned weighted first-order moment of the instance distribution.

While this mechanism effectively captures average signal, it cannot directly model higher-order statistics—such as co-occurrence structures, multimodal feature distributions, or rare-pattern enrichment—which are central to histological heterogeneity and diagnostic accuracy. As a consequence, the architecture implicitly assumes that a slide can be summarized by a single global prototype. This assumption routinely breaks down in heterogeneous whole-slide images, where multiple competing morphologies or subclonal populations may coexist (Zaheer et al., 2017; Lee et al., 2019; Dosovitskiy et al., 2021; Shao et al., 2021).

CLAM partially addresses this issue by using multi-head attention designed to produce class-specific attention maps, but all heads still share a common backbone, limiting their representational diversity. Transformer-based MIL methods in principle can capture richer distributions via self-attention, yet the quadratic complexity of standard attention becomes prohibitive for thousands of patches. TransMIL mitigates this through hierarchical processing with neighbor-restricted (windowed) attentions and cross-scale fusion, reducing effective complexity while retaining contextual information. However, its hierarchical design introduces permutation variance and reduces interpretability—limitations that are problematic for inherently permutation-invariant domains such as cytology.

These observations collectively suggest that effective slide-level analysis requires multiple specialized aggregation mechanisms that can adapt to different morphological patterns within a slide. This naturally points toward architectures that dynamically select or combine diverse processors rather than relying on a single monolithic pooling mechanism.

## 2.2. Mixture-of-Experts for specialized modeling

The Mixture-of-Experts (MoE) framework is a well-established paradigm for scaling model capacity efficiently by employing a set of specialized sub-networks (“experts”) and a learned router that allocates inputs to the most relevant experts. Each expert has its own weight space, enabling specialization without forcing the model to optimize a single compromised solution across partially contradicting objectives (Jacobs et al., 1991; Shazeer et al., 2017). While MoE has been widely adopted in natural language processing, its application in computational pathology has been limited, particularly at the critical aggregation stage.

We propose to address the limitation of unimodal aggregation by introducing a Mixture-of-Experts approach at the slide level, which we denote as Mixture of Aggregators (MoA). Instead of a single aggregator, several permutation-invariant aggregators are trained in

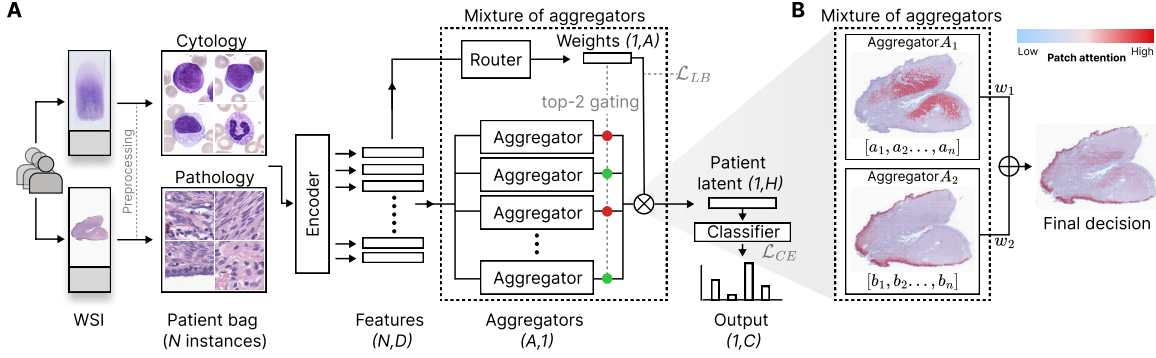


Figure 1: (A) Whole-slide images are patchified and encoded into instance-level embeddings. A router assigns weights and selects the top-2 aggregators, which process the embeddings in parallel. Their weighted outputs are fused into a patient-level latent representation and passed to a classifier for disease prediction. (B) Each aggregator learns to focus on distinct morphological structures within the slide.

parallel, each free to specialize in distinct morphological or domain regimes (for example, immune-rich versus tumor-dominant patterns, rare event sensitivity, or co-occurrence structures). A router produces data-dependent weights (often sparse top- $k$ ) that combine the aggregator summaries for each slide. This preserves a shared feature backbone while adding specialized capacity exactly where heterogeneity is highest—during the instance-to-slide aggregation. In effect, combining multiple invariant summaries via routing enables the model to approximate a richer family of set functions than any single aggregator, while keeping the interface simple (a slide-level vector) and the training recipe close to standard MIL practice (Zaheer et al., 2017; Lee et al., 2019; Shazeer et al., 2017; Lepikhin et al., 2020).

In essence, MoA is not a mere import of MoE into pathology; rather, it re-designs the aggregation stage to reflect the complex nature of histological evidence, bridging the gap between single-aggregator MIL and the heterogeneous reasoning pathologists employ in practice

### 3. Methodology

#### 3.1. MoA: Mixture of aggregators in multiple instance learning

Our framework employs multiple aggregators to capture distinct distributions of instances across different disease types. In our mixture of aggregator experiments, we use two commonly adopted aggregator architectures: Attention-based MIL (ABMIL) and a Transformer-based aggregator. Although many variants of aggregators exist, and in theory one may appear superior to another, recent studies show that their performance is often data-specific and strongly influenced by hyperparameter tuning (Shao et al., 2025). In comparison, we include CLAM, DSMIL and mean pooling (MeanMIL).

For ABMIL, we use the gated attention pooling mechanism of Ilse et al. (Ilse et al., 2018), which learns instance-specific attention weights based on nonlinear gating. The bag representation is then obtained as a weighted sum of the instance embeddings.

As Transformer-based aggregators, we use permutation invariant Transformer proposed by Wagner et al (Wagner et al., 2023) for cytology tasks. Each instance embedding is projected into a 512-dimensional latent space, and a learnable [CLS] token is prepended to the sequence. We use two transformer layers, each consisting of a multi-head self-attention block (with 8 heads) followed by a feedforward network with hidden dimension 1024. For pathology tasks, we use TransMIL (Shao et al., 2021) with same dimensional parameters. [CLS] token is later used as the aggregated representation.

### 3.2. Architecture

We first use an encoder to extract instance-level features (Figure 1). For pathology images, we use UNI (Chen et al., 2024), and for hematologic cytology we use DinoBloom-B (Koch et al., 2024). The extracted features of size (N,D) are passed through a router to compute aggregator weights. The router consists of a projection layer, feature mean pooling and a linear layer. A top-2 softmax gating then selects the most relevant aggregators.

In parallel, the features are passed through all aggregators (Figure 1). The representations from the top-2 selected aggregators are weighted and summed to produce the patient-level latent representation, which is subsequently used for classification.

For optimization, we use standard cross-entropy loss ( $\mathcal{L}_{\text{CE}}$ ) for slide-level classification and add a load-balancing auxiliary loss on the gating network to avoid routing collapse. This loss, following (Fedus et al., 2021), encourages all aggregators to be used more evenly. Let  $A$  be the number of aggregators and  $\mathcal{B}$  the effective batch of  $T$  patient-level bags. During training, we use top-2 routing, i.e., each bag is sent to the two aggregators with the highest router probabilities. For each aggregator  $i \in \{1, \dots, A\}$  we define

$$f_i = \frac{1}{T} \sum_{x \in \mathcal{B}} \mathbf{1}[i \in \text{TopK}(p(x), 2)], \quad P_i = \frac{1}{T} \sum_{x \in \mathcal{B}} p_i(x) \quad (1)$$

where  $p(x) \in \mathbb{R}^A$  is the router’s softmax probability vector for patient  $x$ ,  $p_i(x)$  its  $i$ -th component, and  $\text{TopK}(p(x), 2)$  returns the indices of the two aggregators with highest probability. Thus,  $f_i$  is the fraction of bags actually routed to aggregator  $i$  (hard usage), while  $P_i$  is the average router probability mass assigned to aggregator  $i$  (soft usage). The load-balancing loss is

$$\mathcal{L}_{\text{LB}} = A \sum_{i=1}^A f_i P_i, \quad \mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{lb}} \cdot \mathcal{L}_{\text{LB}}, \quad (2)$$

where  $\lambda_{\text{lb}}$  controls the strength of load balancing.

In addition, to mitigate aggregator collapse and promote diverse aggregator utilization, we perturb the gating logits with independent Gumbel(0,1) noise during training (Shazeer et al., 2017). To gradually transition from exploration to stable specialization, we anneal the softmax temperature: higher temperatures produce smoother, more exploratory distributions across aggregators, while lower temperatures sharpen the distribution and favor more deterministic routing (Nie et al., 2022).

## 4. Experiments

### 4.1. Dataset and preprocessing

We test our model on 2 modalities (cytology and histology), 13 organs/regions and 19 tasks including morphological and immune subtyping. All datasets are publicly available. More details are included in Appendix A.3.

**Cytology:** AML-Hehr (Hehr et al., 2023) and cAItomorph (Dasdelen et al., 2025) are blood smear datasets which include single cell white blood cell images.

**Pathology:** We include fourteen different pathology datasets in our evaluation. These span multiple organ systems, including breast pathology (BCNB (Xu et al., 2021), BRACS (Brancati et al., 2022)), renal cancer (CPTAC-CCRCC), head and neck cancer (CPTAC-HNSC, HANCOCK (Dörrieh et al., 2025)), lung cancer (CPTAC-LSCC), pancreatic cancer (CPTAC-PDA), endometrial cancer (CPTAC-UCEC), cervical cancer (IMP-Cervix (Oliveira et al., 2024)) and other (CPTAC-ALL). ((Ellis et al., 2013; Zhang et al., 2025))). Tasks include biomarker prediction (BCNB), immune class and tumor microenvironment prediction (CPTAC-CCRCC/HNSC/PDA/UCEC, HANCOCK), histological grading (CPTAC/LSCC, HANCOCK, IMP-CERVIX) and tumor site prediction (CPTAC-ALL, HANCOCK).

For all datasets, we fix the test set according to the original publications or the benchmark (Zhang et al., 2025; Vaidya et al., 2025) and report model performances on test set. Within the training split, we perform 5-fold cross-validation.

For cytology datasets, we use the DinoBloom-B hematology feature extractor (Koch et al., 2024). We patchify pathology datasets, using TRIDENT (Zhang et al., 2025; Vaidya et al., 2025) pipeline at  $20\times$  magnification, patch size of  $256 \times 256$ . These patches are then embedded using the UNI feature extractor (Chen et al., 2024).

### 4.2. Evaluation metrics

For multi-class tasks, we report balanced accuracy, while for binary tasks, we report the area under the ROC curve (AUROC).

To analyze instance-level attentions, we extract the attention scores produced by each aggregator. For ABMIL, we directly use the learned attention weights assigned to individual instances. For Transformer-based aggregators, we apply the Attention Rollout method (Abnar and Zuidema, 2020).

Jensen–Shannon divergence (JSD) is utilized to quantify attention distribution differences of aggregators (Lin, 2002).

## 4.3. Results

### 4.3.1. MOA ENHANCES DIAGNOSTIC PREDICTIONS

We evaluate MoA across two modalities (cytology and pathology) and 19 downstream tasks (Table 1). A single, fixed training recipe—selected via the AML-Hehr ablations—is applied consistently to all datasets.

For ABMIL experts, MoA-ABMIL matches or exceeds the single-aggregator ABMIL baseline in almost all settings. Across 19 tasks, MoA-ABMIL underperforms the baseline only once, while providing consistent positive or neutral gains elsewhere. Improvements are marginal on high-performing binary pathology tasks (e.g., BCNB, CPTAC-



Table 1: Comparison of aggregators across datasets.  $\Delta$  indicates relative improvement compared to single aggregator (%). The reported metric is balanced accuracy for multi-class, area under the ROC (AUROC) for binary tasks. **Bold** indicates the best performing model between single vs. mixture of aggregators for the same architecture. Underline highlights the best model across all models.

Dataset (Number of class)	MoA-ABMIL	ABMIL	$\Delta$	MoA-TransMIL	TransMIL	$\Delta$	CLAM-SB	DSMIL	MeanMIL
AML-Hehr (5C)	78.4 $\pm$ 2.2	<b>81.5<math>\pm</math>3.7</b>	-3.8%	<b>81.5<math>\pm</math>1.0</b>	78.6 $\pm$ 2.2	+3.7%	76.7 $\pm$ 6.0	41.7 $\pm$ 2.7	78.3 $\pm$ 2.1
cAltomorph (8C)	<b>52.8<math>\pm</math>1.8</b>	50.9 $\pm$ 1.9	+3.7%	<b>60.1<math>\pm</math>1.1</b>	59.4 $\pm$ 1.9	+1.2%	55.0 $\pm$ 3.0	42.0 $\pm$ 2.5	<u>60.1<math>\pm</math>1.4</u>
BCNB/ER (2C)	<b>91.3<math>\pm</math>0.2</b>	91.2 $\pm$ 0.4	+0.1%	<b>88.4<math>\pm</math>0.5</b>	85.7 $\pm$ 3.3	+3.1%	90.8 $\pm$ 0.7	90.9 $\pm$ 0.5	<u>91.7<math>\pm</math>0.4</u>
BCNB/HER2 (2C)	<b>84.0<math>\pm</math>0.7</b>	84.0 $\pm$ 0.5	0.0%	<b>81.2<math>\pm</math>2.2</b>	69.8 $\pm$ 3.5	+16.4%	83.1 $\pm$ 2.2	81.5 $\pm$ 1.3	<u>84.1<math>\pm</math>1.1</u>
BCNB/PR (2C)	<b>88.3<math>\pm</math>0.4</b>	88.2 $\pm$ 0.6	+0.1%	<b>84.7<math>\pm</math>0.7</b>	78.2 $\pm$ 3.5	+8.3%	87.1 $\pm$ 1.2	86.4 $\pm$ 0.4	<u>89.0<math>\pm</math>0.5</u>
BRACS (7C)	<b>34.6<math>\pm</math>1.9</b>	34.4 $\pm$ 1.5	+0.6%	<b>29.3<math>\pm</math>1.6</b>	26.5 $\pm$ 1.5	+10.6%	32.8 $\pm$ 3.3	27.2 $\pm$ 2.5	26.9 $\pm$ 1.7
CPTAC-ALL (10C)	<b>96.1<math>\pm</math>0.3</b>	95.5 $\pm$ 0.2	+0.6%	<b>96.3<math>\pm</math>0.6</b>	95.5 $\pm$ 1.0	+0.8%	96.4 $\pm$ 0.7	96.5 $\pm$ 0.4	<u>96.8<math>\pm</math>0.6</u>
CPTAC-CCRCC (3C)	<b>45.4<math>\pm</math>7.7</b>	43.5 $\pm$ 4.3	+4.4%	<b>47.2<math>\pm</math>4.1</b>	45.2 $\pm$ 3.7	+4.4%	<u>47.7<math>\pm</math>4.0</u>	45.4 $\pm$ 3.7	45.9 $\pm$ 4.7
CPTAC-HNSC (3C)	<b>35.1<math>\pm</math>5.7</b>	33.0 $\pm$ 4.6	+6.4%	<b>31.6<math>\pm</math>3.6</b>	27.9 $\pm$ 5.8	+13.3%	35.1 $\pm$ 3.9	30.2 $\pm$ 3.0	34.5 $\pm$ 5.6
CPTAC-LSCC (2C)	<b>69.8<math>\pm</math>2.2</b>	67.1 $\pm$ 3.5	+4.0%	<b>63.7<math>\pm</math>5.2</b>	60.0 $\pm$ 9.4	+6.2%	65.1 $\pm$ 2.3	65.0 $\pm$ 2.8	60.2 $\pm$ 3.4
CPTAC-PDA (3C)	<b>39.3<math>\pm</math>7.3</b>	35.1 $\pm$ 3.0	+11.9%	<b>41.3<math>\pm</math>4.1</b>	32.9 $\pm$ 7.0	+25.5%	40.0 $\pm$ 6.1	36.2 $\pm$ 3.1	40.1 $\pm$ 2.6
CPTAC-UCEC (3C)	<b>43.2<math>\pm</math>2.8</b>	36.3 $\pm$ 5.6	+19.0%	<b>44.9<math>\pm</math>7.3</b>	29.7 $\pm$ 7.9	+51.2%	37.0 $\pm$ 9.4	28.7 $\pm$ 7.8	33.5 $\pm$ 10.1
HANCOCK/K-SCC grading (2C)	<b>73.8<math>\pm</math>1.3</b>	71.4 $\pm$ 5.8	+3.4%	<b>73.6<math>\pm</math>2.9</b>	60.8 $\pm$ 6.6	+21.1%	67.6 $\pm$ 2.3	54.7 $\pm$ 9.2	70.8 $\pm$ 4.9
HANCOCK/NK-SCC grading (2C)	<b>67.0<math>\pm</math>5.8</b>	62.0 $\pm$ 10.8	+8.1%	<b>61.0<math>\pm</math>10.2</b>	48.0 $\pm$ 9.7	+27.1%	62.5 $\pm$ 8.2	46.0 $\pm$ 15.5	53.0 $\pm$ 12.2
HANCOCK/perineural invasion (2C)	<b>79.8<math>\pm</math>1.1</b>	76.9 $\pm$ 0.7	+3.8%	<b>75.5<math>\pm</math>3.2</b>	63.9 $\pm$ 5.6	+18.1%	75.6 $\pm$ 2.9	63.0 $\pm$ 7.8	76.1 $\pm$ 0.7
HANCOCK/metastasis (2C)	<b>74.8<math>\pm</math>1.3</b>	71.4 $\pm$ 1.7	+4.8%	<b>64.7<math>\pm</math>3.6</b>	63.2 $\pm$ 5.5	+2.4%	67.1 $\pm$ 6.5	62.6 $\pm$ 6.8	73.3 $\pm$ 3.6
HANCOCK/tumor site (4C)	<b>74.1<math>\pm</math>3.2</b>	68.5 $\pm$ 1.9	+8.2%	<b>71.8<math>\pm</math>3.2</b>	66.7 $\pm$ 1.0	+7.6%	71.3 $\pm$ 2.0	60.4 $\pm$ 2.2	71.1 $\pm$ 2.4
HANCOCK/vascular invasion (2C)	<b>55.3<math>\pm</math>6.8</b>	51.6 $\pm$ 7.6	+7.2%	<b>66.8<math>\pm</math>3.8</b>	59.9 $\pm$ 6.0	+11.5%	62.1 $\pm$ 6.4	52.8 $\pm$ 8.0	55.3 $\pm$ 9.9
IMP-Cervix (3C)	<b>46.6<math>\pm</math>2.8</b>	45.0 $\pm$ 3.6	+3.6%	<b>57.0<math>\pm</math>1.9</b>	52.9 $\pm$ 4.5	+7.7%	<u>61.0<math>\pm</math>4.6</u>	47.0 $\pm$ 6.7	48.9 $\pm$ 1.9
<b>Average</b>	<b>64.7</b>	<b>62.5</b>	<b>+4.5%</b>	<b>64.2</b>	<b>58.1</b>	<b>+12.6%</b>	<b>63.9</b>	<b>55.7</b>	<b>62.6</b>

ALL;  $\Delta \approx 0.1\text{--}0.8\%$ ), and become more pronounced on harder multi-class cohorts with lower baselines. For example, MoA-ABMIL improves balanced accuracy on CPTAC-HNSC (+6.4%), CPTAC-LSCC (+4.0%), CPTAC-PDA (+11.9%) and CPTAC-UCEC (+19.0%), and achieves the best overall performance on several HANCOCK grading and invasion tasks. On average, MoA-ABMIL yields a +4.5% relative improvement over the single-aggregator ABMIL baseline across all datasets.

For mixture of Transformer experts, the effect is even stronger. MoA-TransMIL strictly dominates the single TransMIL baseline on every dataset, with an average relative gain of +12.6% (Table 1). Improvements are modest on already saturated tasks (e.g., CPTAC-ALL: +0.8%), but become substantial on more challenging settings: BRACS (+10.6%), BCNB/HER2 (+16.4%), BCNB/PR (+8.3%), CPTAC-PDA (+25.5%), and especially CPTAC-UCEC (+51.2%). Similar trends are observed across the HANCOCK, where MoA-TransMIL delivers large gains for K-SCC grading (+21.1%), NK-SCC grading (+27.1%), perineural invasion (+18.1%), vascular invasion (+11.5%). In cytology, MoA-TransMIL also improves over the baseline for AML-Hehr (+4.5%) and cAltomorph (+1.2%), confirming that the benefits of mixture-of-aggregators transfer to both smear and tissue-based tasks.

Compared to alternative MIL baselines, MoA is competitive or superior on most datasets. Although the simplest, MeanMIL attains the best performance on 4/19 tasks (BCNB and CPTC-ALL), while CLAM-SB slightly outperforms other models on CPTAC-CCRCC and IMP-Cervix. Nevertheless, either MoA-ABMIL or MoA-TransMIL is the best or tied-best model on the majority of tasks (13/19). These results indicate that mixing aggregators—regardless of the choice of backbone—yields a robust improvement over both single-aggregator variants and established MIL baselines.

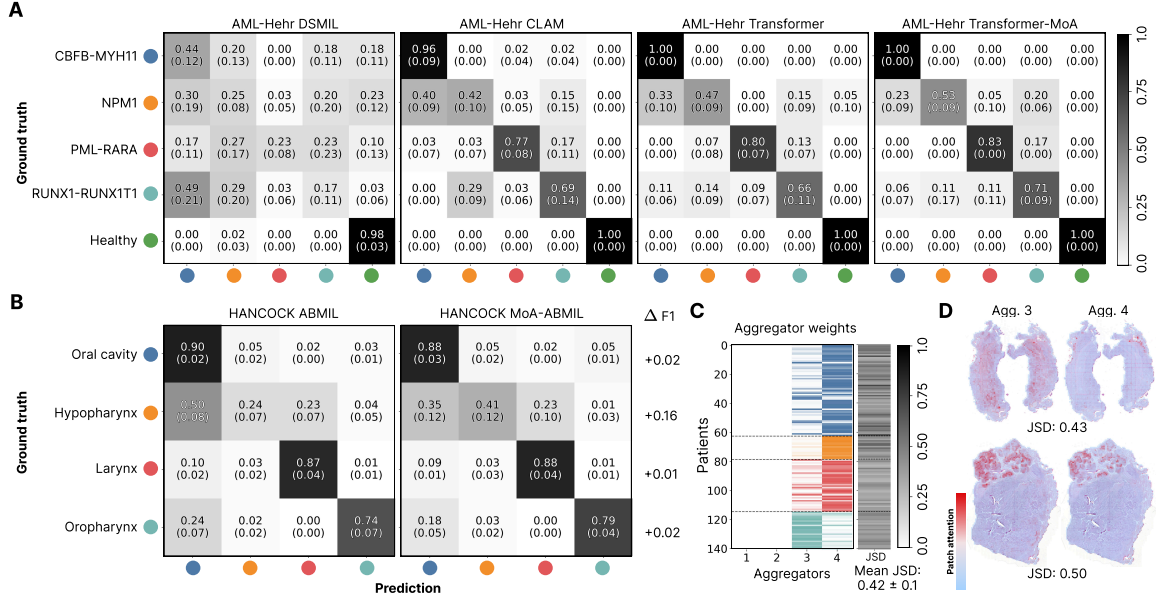


Figure 2: (A) Confusion matrices for AML-Hehr dataset. (B) MoA improves class specific F1 scores. (C) Aggregator weight distributions across patients reveal class-specific specialization. (D) Attentions given by aggregators differ for patients and quantified by Jensen-Shannon divergence (JSD).

The additional benefit MoA on cytology and histology tasks are shown in Figure 2 with confusion matrices. Beyond overall performance, our method enhances class-level sensitivity and reduces confusion between malignant and non-malignant categories in the AML-Hehr dataset (Figure 2A). Our method consistently improve performance of all classes in HANCOCK dataset (Figure 2B). Importantly, the aggregator weight distributions reveal distinct specializations (Figure 2C). In the primary tumor-site identification task, Aggregator 3 predominantly contributes to oropharynx cases, while Aggregator 4 is more active in other tumor regions. Jensen-Shannon divergence analysis shows that the aggregators attain different attention distributions over patches (Figure 2C, D), with a mean of  $0.42 \pm 0.10$ .

#### 4.3.2. AGGREGATORS CAPTURE DISTINCT ATTENTION PATTERNS

To further assess aggregator specialization and determine whether they capture different distributions within a bag, we conducted a patient-wise analysis and computed instance-level attentions (single-cell images) from each aggregator (Figure 3). We focus on the AML-Hehr hematologic cytology dataset because single-cell contributions to disease are easier to assess and several AML subtypes exhibit pathognomonic morphologic findings. We present three representative patients from different AML subtypes. The upper panel shows the attention distributions generated by the aggregators along with their respective contribution weights.

Patient YST carries a PML::RARA fusion, also known as acute promyelocytic leukemia (APL), a distinct subtype of myeloid leukemia that requires rapid diagnosis and treatment. The hallmark of APL is the presence of promyelocytes. In patient YST, the dominant



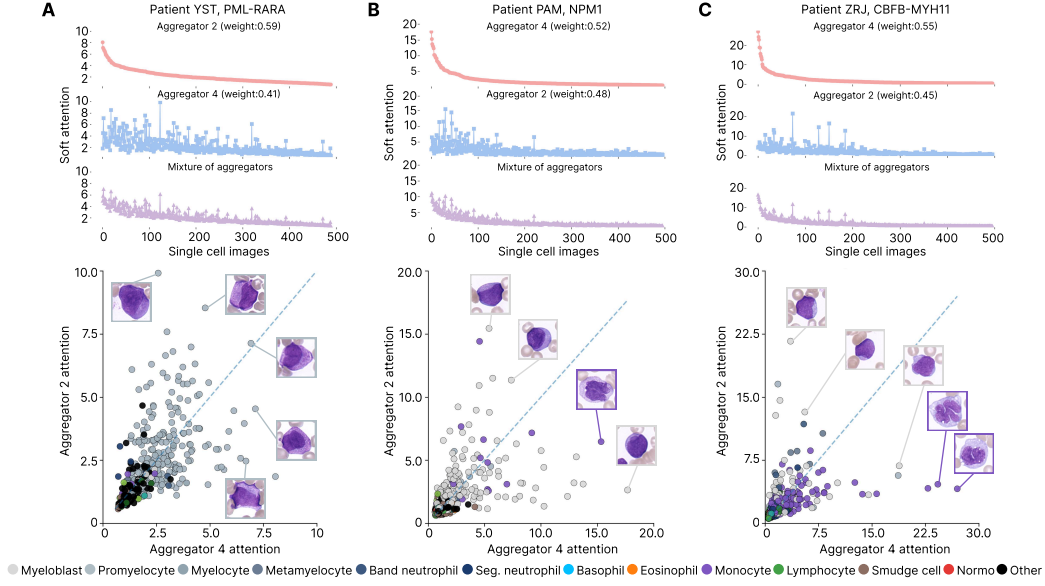


Figure 3: Different aggregators capture complementary diagnostic patterns. For the AML-Hehr dataset, we show patient-level attention analyses for (A) AML with PML-RARA, (B) AML with NPM1, and (C) AML with CBFB-MYH11. Top panels display single-cell attention distributions for the top-2 contributing aggregators and their mixture weights (red: dominant, blue: secondary, purple: mixture). Bottom panels compare their attentions, illustrating how each focuses on distinct cell subsets. This complementarity enables the model to capture subtype-specific morphological patterns. Attention scores are reported in  $10^{-3}$  units.

aggregator (Aggregator 2) assigns high attention to several promyelocytes, while the second aggregator highlights additional promyelocytes initially overlooked by the first (Figure 3A).

In patient PAM, the aggregators complement each other, with both capturing myeloblasts, which are essential for AML diagnosis (Figure 3B).

Patient ZRJ harbors a CBFB::MYH11 fusion, a genetic abnormality associated with AML characterized by monocytic and granulocytic differentiation. In this case, the first aggregator (Aggregator 4) identifies monocytic cells as disease-specific instances (Figure 3C). The second aggregator complements by assigning high attention to myeloblasts and granulocytic cells at different maturation stages.

#### 4.4. Ablation Study

We select a single, fixed MoA configuration via ablations on AML-Hehr (Table 2) and apply to the rest of the dataset. Compared to the single-aggregator baseline (78.6 balanced accuracy), mixtures only help when routing is lightly regularized and stochastic. Performance is best with an intermediate number of aggregators and mild load balancing: a router with four aggregators,  $\lambda_{lb} = 0.01$ , and Gumbel noise enabled achieves 81.5 balanced accuracy on AML-Hehr (+3.7% vs. base). Stronger regularization ( $\lambda_{lb} = 0.10$ ), disabling Gumbel

Table 2: Selected router configurations on AML-Hehr. *Base* = baseline (single aggregator, no router); *MLP* = multi-layer perceptron. Second line shows number of aggregators used. Third line shows load-balancing loss coefficient  $\lambda_{lb}$ ; fourth line indicates whether Gumbel routing is used (True/False). **Bold** indicates the best-performing configuration.

<i>Router architecture:</i>	Base	<b>Linear</b>	MLP	Linear	Linear	Linear	Linear
<i>Number of aggregators:</i>	–	<b>4</b>	4	2	6	4	4
$\lambda_{lb}$ :	–	<b>0.01</b>	0.01	0.01	0.01	0.10	0.01
<i>Gumbel noise:</i>	–	<b>T</b>	T	T	T	T	F
<b>Balanced Acc</b>	78.6	<b>81.5</b>	76.1	76.8	79.3	80.0	78.1
<b>% vs Base</b>	–	<b>+3.7</b>	-3.2	-2.3	+0.9	+1.8	-0.6

noise, or changing the number of aggregators (2 or 6) reduces performance. We therefore adopt this 4-aggregator, Gumbel +  $\lambda_{lb} = 0.01$  configuration as the fixed training recipe for all subsequent experiments, where it generalizes well across organs and modalities (full ablations in Appendix B).

#### 4.5. Limitations

We acknowledge several limitations of our study. First, as shown in prior work, no single MIL architecture uniformly outperforms all others; performance is strongly dataset dependent. In our experiments, mixtures of aggregators improve over their single-aggregator counterparts, but they may not always achieve the best performance compared to alternative MIL architectures. Finally, although we used a fixed training recipe across all datasets, additional hyperparameter tuning may be necessary to benefit of MoA in other settings.

## 5. Conclusion

We propose Mixture of Aggregators (MoA), a framework that employs multiple aggregators for multiple instance learning to better capture diverse distributions in heterogeneous datasets. Our approach improves diagnostic performance in both pathology and hematologic cytology. Within this framework, aggregators specialize in different disease types, and each specialized aggregator provides distinct, clinically relevant attention distributions over the instances. These complementary attention patterns enhance diagnostic accuracy when combined.

## References

- Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. URL <https://arxiv.org/abs/2005.00928>.
- M. Bilal, R. J. Chen, J. van der Laak, T. de Bruyn, G. Litjens, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *The Lancet Digital Health*, 3(e8):e763–e772, 2021. doi: 10.1016/S2589-7500(21)00133-3.

- BiOPTImus. H-optimus-1, 2025. URL <https://huggingface.co/biOPTImus/H-optimus-1>.
- Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierto, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022: baac093, 2022.
- Wouter Bulten, Maschenka Balkenhol, Jean-Joël Awoumou Belinga, Américo Brilhante, Ash Çakır, Lars Egevad, Martin Eklund, Xavier Farré, Katerina Geronatsiou, Vincent Molinié, Guilherme Pereira, Paromita Roy, Günter Saile, Paulo Salles, Ewout Schaafsma, Joëlle Tschui, Anne-Marie Vos, ISUP Pathology Imagebase Expert Panel, Hester van Boven, Robert Vink, Jeroen van der Laak, Christina Hulsbergen-van der Kaa, Geert Litjens, Brett Delahunt, Hemamali Samarasinghe, David J. Grignon, Andrew J. Evans, Daniel M. Berney, Chin-Chen Pan, Glen Kristiansen, James G. Kench, Jon Oxley, Katia R.M. Leite, Jesse K. McKenney, Peter A. Humphrey, Samson W. Fine, Toyonori Tsuzuki, Murali Varma, Ming Zhou, Eva Comperat, David G. Bostwick, Kenneth A. Iczkowski, Cristina Magi-Galluzzi, John R. Srigley, Hiroyuki Takahashi, and Theo van der Kwast. Artificial intelligence assistance significantly improves gleason grading of prostate biopsies by pathologists. *Modern Pathology*, 34(3):660–671, 2021. ISSN 0893-3952. doi: <https://doi.org/10.1038/s41379-020-0640-y>. URL <https://www.sciencedirect.com/science/article/pii/S0893395222006652>.
- Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. doi: 10.1038/s41591-019-0508-1.
- Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.
- Muhammed Furkan Dasdelen, Ivan Kukuljan, Peter Lienemann, Fatih Ozlugedik, Ario Sadafi, Matthias Hehr, Karsten Spiekermann, Christian Pohlkamp, and Carsten Marr. Transformer-based hematological malignancy prediction from peripheral blood smears in a real-world cohort. *arXiv preprint arXiv:2509.20402*, 2025.
- Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, Drew F. K. Williamson, Bowen Chen, Cristina Almagro-Perez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Daisuke Komura, Akihiro Kawabe, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, and Faisal Mahmood. Multimodal whole slide foundation model for pathology, 2024. URL <https://arxiv.org/abs/2411.19666>.
- Marion Dörrich, Matthias Balk, Tatjana Heusinger, Sandra Beyer, Hamed Mirbagheri, David J Fischer, Hassan Kanso, Christian Matek, Arndt Hartmann, Heinrich Iro, et al. A multimodal dataset for precision oncology in head and neck cancer. *Nature Communications*, 16(1):7163, 2025.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- A. Dy, N. J. Nguyen, J. Meyer, M. Dawe, W. Shi, D. Androustos, A. Fyles, F. F. Liu, S. Done, and A. Khademi. Ai improves accuracy, agreement and efficiency of pathologists for ki67 assessments

- in breast cancer. *Scientific Reports*, 14(1):1283, 2024. doi: 10.1038/s41598-024-51723-2. URL <https://doi.org/10.1038/s41598-024-51723-2>.
- MJ Ellis, M Gillette, SA Carr, AG Paulovich, RD Smith, KK Rodland, RR Townsend, C Kinsinger, M Mesri, H Rodriguez, et al. Clinical proteomic tumor analysis consortium (cptac): Connecting genomic alterations to cancer biology with proteomics: the nci clinical proteomic tumor analysis consortium. *Cancer Discov*, 3:1108–1112, 2013.
- William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*, 2021. URL <https://arxiv.org/abs/2101.03961>.
- Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024.
- Matthias Hehr, Ario Sadafi, Christian Matek, Peter Lienemann, Christian Pohlkamp, Torsten Haferlach, Karsten Spiekermann, and Carsten Marr. Explainable ai identifies diagnostic cells of genetic aml subtypes. *PLOS Digital Health*, 2(3):e0000187, 2023.
- Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 2018. URL <https://proceedings.mlr.press/v80/ilse18a.html>.
- Maximilian Ilse, Jakub Tomczak, and Max Welling. *Deep multiple instance learning for digital histopathology*, pages 521–546. 01 2020. ISBN 9780128161760. doi: 10.1016/B978-0-12-816176-0.00027-2.
- Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.
- Andrew Janowczyk, Rui Zuo, Heather Gilmore, Michael Feldman, and Anant Madabhushi. Histoqc: An open-source quality control tool for digital pathology slides. *JCO Clinical Cancer Informatics*, 3:1–7, April 2019. doi: 10.1200/CCI.18.00157.
- Hassan Keshvarikhajasteh. Quantitative evaluation of multiple instance learning reliability for wsis classification, 2025. URL <https://arxiv.org/abs/2409.11110>.
- Valentin Koch, Sophia J Wagner, Salome Kazemina, Ece Sancar, Matthias Hehr, Julia A Schnabel, Tingying Peng, and Carsten Marr. Dinobloom: a foundation model for generalizable cell embeddings in hematology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 520–530. Springer, 2024.
- Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.
- Dmitry Lepikhin, HyounJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. URL <https://arxiv.org/abs/2006.16668>.

- Bin Li, Yin Li, and Kevin W Eliceiri. 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*, pages 14318–14328, 2021.
- Yan-Jun Li, Hsin-Hung Chou, Peng-Chan Lin, Meng-Ru Shen, Sun-Yuan Hsieh, et al. A novel deep learning-based algorithm combining histopathological features with tissue areas to predict colorectal cancer survival from whole-slide images. *Journal of Translational Medicine*, 21(1):731, 2023. doi: 10.1186/s12967-023-04530-8.
- Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 2002.
- Ming Y Lu, Drew F K Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.
- Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, Georg Gerber, Anil V Parwani, and Faisal Mahmood. Towards a visual-language foundation model for computational pathology, 2023. URL <https://arxiv.org/abs/2307.12914>.
- Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.
- National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium clear cell renal cell carcinoma collection (cptac-ccrc) (version 14). The Cancer Imaging Archive, 2018a. URL <https://doi.org/10.7937/K9/TCIA.2018.OBLAMN27>. [Dataset].
- National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium head and neck squamous cell carcinoma collection (cptac-hnsc) (version 19). The Cancer Imaging Archive, 2018b. URL <https://doi.org/10.7937/K9/TCIA.2018.UW45NH81>. [Dataset].
- National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium lung squamous cell carcinoma collection (cptac-lsc) (version 15). The Cancer Imaging Archive, 2018c. URL <https://doi.org/10.7937/K9/TCIA.2018.6EMUB5L2>. [Dataset].
- National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium uterine corpus endometrial carcinoma collection (cptac-ucec) (version 13). The Cancer Imaging Archive, 2019. URL <https://doi.org/10.7937/K9/TCIA.2018.3R3JUISW>. [Dataset].
- Xiaonan Nie, Shijie Cao, Xupeng Miao, Lingxiao Ma, Jilong Xue, Youshan Miao, Zichao Yang, Zhi Yang, and Bin CUI. Dense-to-sparse gate for mixture-of-experts, 2022. URL [https://openreview.net/forum?id=\\_4D8IVs7y08](https://openreview.net/forum?id=_4D8IVs7y08).
- Sara P. Oliveira, Diana Montezuma, Ana Moreira, Domingos Oliveira, Pedro C. Neto, Ana Monteiro, João Monteiro, Liliana Ribeiro, Sofia Gonçalves, Isabel M. Pinto, and Jaime S. Cardoso. Imp whole-slide images of cervical samples 2024, 2024. URL <https://doi.org/10.25747/9028-C140>. Data set.

- Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts, 2021. URL <https://arxiv.org/abs/2106.05974>.
- Ario Sadafi, Asya Makhro, Anna Bogdanova, Nassir Navab, Tingying Peng, Shadi Albarqouni, and Carsten Marr. Attention based multiple instance learning for classification of blood cell disorders. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 246–256. Springer, 2020.
- Daniel Shao, Richard J. Chen, Andrew H. Song, Joel Runevic, Ming Y. Lu, Tong Ding, and Faisal Mahmood. Do multiple instance learning models transfer?, 2025. URL <https://arxiv.org/abs/2506.09022>.
- Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017. URL <https://openreview.net/forum?id=BlckMDqIg>.
- David F. Steiner, Robert MacDonald, Yun Liu, Peter Truszkowski, Jason D. Hipp, Christopher Gammage, Florence Thng, Lily Peng, and Martin C. Stumpe. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *The American Journal of Surgical Pathology*, 42(12):1636–1646, 2018. ISSN 0147-5185. doi: 10.1097/PAS.0000000000001151.
- Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H Song, Tong Ding, Sophia J Wagner, Ming Y Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, et al. Molecular-driven foundation model for oncologic pathology. *arXiv preprint arXiv:2501.16652*, 2025.
- Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Ellen Yang, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C.H. Lee, Jan H. Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Hannah Wen, Juan A. Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David S. Klimstra, Brandon Rothrock, Siqi Liu, and Thomas J. Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, 2024. doi: 10.1038/s41591-024-03141-0.
- Sophia J Wagner, Daniel Reisenbüchler, Nicholas P West, Jan Moritz Niehues, Jiefu Zhu, Sebastian Foersch, Gregory Patrick Veldhuizen, Philip Quirke, Heike I Grabsch, Piet A van den Brandt, et al. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer cell*, 41(9):1650–1661, 2023.
- Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in oncology*, 11:759007, 2021.
- Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems 30*, pages 3391–3401. Curran Associates, Inc., 2017.



Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750*, 2025.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

## Appendix A. Supplementary methods

### A.1. Aggregator details

For ABMIL, we use the gated attention pooling mechanism introduced by Ilse et al (Ilse et al., 2018). Given a bag of  $n$  instance embeddings  $h_{i=1}^n$ , the attention weight for each instance is computed as

$$a_i = \frac{\exp\{w^\top [\tanh(Vh_i) \odot \sigma(Uh_i)]\}}{\sum_{j=1}^n \exp\{w^\top [\tanh(Vh_j) \odot \sigma(Uh_j)]\}}, \quad (3)$$

where  $V, U \in \mathbb{R}^{l \times d}$  and  $w \in \mathbb{R}^l$  are learnable parameters,  $\tanh(\cdot)$  and  $\sigma(\cdot)$  denote the hyperbolic tangent and sigmoid activations, and  $\odot$  is the element-wise product. The bag-level representation is then obtained as a weighted sum of instances.

$$z = \sum_{i=1}^n a_i h_i. \quad (4)$$

For Transformer architecture, we use recipe by (Wagner et al., 2023). Self attention is defined as:

$$SA(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad (5)$$

where queries  $Q \in \mathbb{R}^{n \times d_k}$ , keys  $K \in \mathbb{R}^{n \times d_k}$ , and values  $V \in \mathbb{R}^{n \times d_v}$  are obtained from input embeddings  $x$  via

$$Q = W_Q x, \quad K = W_K x, \quad V = W_V x, \quad (6)$$

with learnable weights  $W_Q \in \mathbb{R}^{d \times d_k}$ ,  $W_K \in \mathbb{R}^{d \times d_k}$ , and  $W_V \in \mathbb{R}^{d \times d_v}$ .

### A.2. Training details

For model training, we use a fixed learning rate of  $5 \times 10^{-5}$  with the AdamW optimizer, update gradients every 16 patients, and train for 150 epochs with early stopping based on the validation loss. We employ four aggregators and adopt a staged training schedule: during the first three epochs, all four aggregators are used equally, after which we gradually reduce the active expert count to two. The goal of this warm-up phase is to ensure that all aggregators acquire a minimal, shared understanding of the task before the router begins to decide which aggregators are most relevant for each sample. Enforcing top-2 routing encourages the aggregators to specialize rather than collapsing into a simple ensemble.

We used single H100 80GB GPU for running the experiments.

### A.3. Dataset details

**AML-Hehr** (Hehr et al., 2023) includes 189 patients from four acute myeloid leukemia (AML) genetic subtypes (PML-RARA fusion, NPM1-mutation, CBFB-MYH11 fusion and RUNX1-RUNX1T1 fusion) and healthy controls. Each patient has an average of  $430 \pm 107$  single-cell white blood cell images. We hold out 43 patients as the test set.

**cAItomorph** data (Dasdelen et al., 2025) includes 2,043 patients spanning seven hematologic conditions—acute myeloid leukemia (AML), myelodysplastic syndromes (MDS), myeloproliferative neoplasms (MPN), MDS/MPN overlap syndromes, lymphoma, plasma cell neoplasms, and reactive changes—along with a healthy cohort. The number of white blood cell images per patient ranges from 55 to 500, with an average of  $488 \pm 55$  cells. A total of 409 patients are held out for testing. This dataset is particularly challenging due to substantial inter-class heterogeneity and intra-class overlap.

**BCNB** (Xu et al., 2021) includes 1058 core needle biopsy slides from early breast cancer patients. It includes binary prediction tasks of ER, PR and HER2 status.

**BRACS** (Brancati et al., 2022) consist of 547 breast tissue biopsy slides from 189 patients. It includes 7 classes: normal, pathological benign, usual ductal hyperplasia, flat epithelial atypia, qtypical ductal hyperplasia, ductal carcinoma in situ and invasive carcinoma.

**CPTAC-CCRCC** (National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC), 2018a) has 103 slides with clear cell renal cell carcinoma. We predict the immune class of patients: low, medium, high.

**CPTAC-HNSCC** (National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC), 2018b) has 107 slides with head and neck cancer. Immune class prediction is made (low, medium, high).

**CPTAC-LSCC** (National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC), 2018c) includes 104 slides with lung squamous cell carcinoma. Histologic grade of the patients are predicted (well differentiated vs moderately differentiated)

**CPTAC-UCEC** (National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC), 2019) has 94 slides with endometrial carcinoma. Immune class prediction is made (low, medium, high).

**HANCOCK** (Dörrich et al., 2025) is a multimodal dataset of 763 head and neck cancer patients. We include 6 different task from this dataset. 5 of them are binary including: Keratinizing squamous cell carcinoma grading ( $n=383$ ), non keratinizing squamous cell carcinoma grading ( $n=74$ ), lymphovascular invasion ( $n=697$ ), perineural invasion ( $n=697$ ) and primary vs. metastasis tumor ( $n=676$ ). Prediction of primary tumor site includes 696 cases from 4 different location (oral cavity, larynx, oropharynx, hypopharynx)

**IMP-Cervix** (Oliveira et al., 2024) includes 5333 samples from cervical biopsy. Cervical cancer grade is predicted: non-neoplastic, low-grade, high-grade.

For train-test splits, we follow PathoBench benchmark (Zhang et al., 2025). We held out the test split and apply 5 cross-validation within training set.

## Appendix B. Supplementary findings

Table 3: Full ablation of different configurations on AML-Hehr data. *Base* = baseline (single aggregator, no router); *Lin.* = Linear; *MLP* = multi-layer perceptron. Second line shows number of experts used. Third line for load balancing loss coeff  $\lambda_B$ ; fourth line Gumbel routing (True/False). **Bold** indicates the best-performing configuration

<i>router-arch:</i>	Base	MLP	Lin	Lin	MLP	MLP	Lin	Lin	MLP	MLP	Lin	Lin	MLP	Lin	MLP	MLP	Lin	Lin	MLP	MLP	<b>Lin</b>	Lin	MLP	MLP	Lin
<i>#Experts:</i>	–	2	2	2	2	4	4	4	4	6	6	6	6	2	2	2	2	4	4	4	<b>4</b>	6	6	6	6
$\lambda_B$ :	–	0.01	0.01	0.10	0.10	0.01	0.01	0.10	0.10	0.01	0.01	0.10	0.10	0.10	0.01	0.10	0.01	0.10	0.01	0.10	<b>0.01</b>	0.10	0.01	0.10	0.01
<i>Gumbel:</i>	–	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	<b>T</b>	T	T	T	T
<b>Balanced Acc</b>	78.6	79.1	78.5	78.5	79.1	74.9	78.1	75.8	78.5	79.1	78.1	77.6	72.2	76.8	79.6	77.8	76.8	80.0	76.1	76.3	<b>81.5</b>	74.1	78.0	78.2	79.3
<b>% vs Base</b>	–	+0.6	-0.1	-0.1	+0.6	-4.7	-0.6	-3.6	-0.1	+0.6	-0.6	-1.3	-8.1	-2.3	+1.3	-1.0	-2.3	+1.8	-3.2	-2.9	<b>+3.7</b>	-5.7	-0.8	-0.5	+0.9

Table 4: Inference times of different MIL architectures on the HANCOCK dataset. MoA increases inference time by approximately  $2.5\times$  compared to a single-aggregator model.

Architecture	Total time [s]	Time / sample [ms]
ABMIL	0.0518	0.3677
MoA-ABMIL	0.1483	1.0515
TransMIL	0.4617	3.2744
MoA-TransMIL	1.1004	7.8046
DSMIL	0.1009	0.7158
CLAM-SB	0.1976	1.4014
Mean	0.0245	0.1734