# OGPCMIL: OMICS-GUIDED PROTOTYPE CALIBRATION MULTIMODAL MULTIPLE INSTANCE LEARNING FOR FEW-SHOT WHOLE SLIDE IMAGE SUBTYPING OF RARE CANCERS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Accurate subtype classification of rare cancers is essential for individualized clinical treatment. However, rare cancers suffer from limited cases and scarce annotations, so unimodal pathological data fail to capture subtype-specific features and stably distinguish fine-grained subtypes. Multimodal learning integrating whole slide images (WSIs) and genomic data provides a feasible solution, yet existing multimodal and multi-instance learning methods still have key limitations for few-shot rare cancer subtyping: (1) Most methods directly establish dense cross-modal interactions between high-dimensional omics features and numerous image patches through global fusion. Such operations easily cause overfitting under limited labeled data and fail to capture stable subtype distinctions. (2) Genomic profiles reflect the overall molecular state of patients, while subtype-specific pathological features only exist in local WSI regions. Current approaches cannot effectively connect global omics priors with local patch features, resulting in underutilization of cross-modal complementary information. To address these issues, we propose OGPCMIL, an Omics-Guided Prototype Calibration Multimodal Multiple Instance Learning framework for rare cancer subtyping. We first adopt a low-rank prototype calibration mechanism to transform high-dimensional omics information into subtle and reliable subtype prototype adjustments, effectively avoiding overfitting in few-shot cross-modal fusion. Furthermore, we design a class-conditional attention aggregation mechanism, which delivers global omics priors to local patches guided by calibrated prototypes, establishing precise associations between omics signatures and subtype-related pathological regions. Extensive 1-shot, 5-shot and 10-shot experiments on three rare cancer datasets show that OGPCMIL outperforms all compared baselines in 18 of 27 settings. Ablation studies validate the effectiveness of our core modules.

## 1 INTRODUCTION

Rare cancers collectively account for approximately 20%–25% of all malignant tumors, yet their molecularly distinct subtypes often remain unresolved because annotated cases are extremely scarce (Xue & Lyu, 2024; DeSantis et al., 2017; van der Graaf et al., 2025). Accurate subtype classification is essential for individualized treatment, as subtypes within the same cancer can have markedly different prognoses and therapeutic responses (Zhang & Yue, 2016). Whole-slide images (WSIs) alone struggle to capture the fine-grained morphological signatures required for subtype discrimination, especially when only a handful of labeled slides are available. Multimodal learning that couples WSIs with genomic profiles offers a natural remedy: tissue morphology and molecular phenotypes are inherently complementary, and recent work has even shown that deep models can predict RNA-seq expression directly from histology images, underscoring the bidirectional association between the two modalities (Schmauch et al., 2020; Pizurica et al., 2024). In the few-shot regime typical of rare cancers, however, simply fusing these two modalities can backfire. Cross-modal interactions are high-dimensional, and MIL provides only weak slide-level supervision, with the discriminative patches left unannotated, so overfitting becomes highly likely. This paper identifies and tackles two obstacles that prevent effective multimodal few-shot subtyping.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The first obstacle is cross-modal overfitting under sample scarcity. Standard multimodal MIL pipelines (Chen et al., 2020; 2022) encourage dense interactions between high-dimensional genomic vectors and thousands of patch embeddings. When training cases are extremely limited, these interactions have far more degrees of freedom than the data can constrain, causing the model to latch onto spurious correlations that separate the few training samples. Prototype-based methods simplify the decision boundary, but naively injecting genomics into prototype construction re-introduces high-capacity fusion pathways and loses that benefit.

The second, more subtle obstacle is the misalignment between global genomic priors and localized pathological evidence. A patient's transcriptomic profile summarizes the tumor's global molecular state, whereas subtype hallmarks are often confined to tiny, scattered WSI regions (Lee, 2023). Existing fusion schemes treat genomics as a global descriptor to be concatenated or gated with slide-level representations. They provide no mechanism to propagate that signal to the patch level, so the model cannot learn where to look for subtype-relevant morphology, and the complementary value of the two modalities remains unrealized.

We address these two obstacles with **OGPCMIL** (Omics-Guided Prototype Calibration Multimodal MIL), a framework built around a single design principle, in which genomics serves as a low-capacity calibrating prior in the few-shot setting. Two modules implement this principle.

First, the **Omics-Guided Low-Rank Prototype Calibration** module. Instead of directly fusing omics features with patches, we use the patient's gene expression profile to generate a low-rank correction to a set of learnable subtype prototypes. The correction is factorized along a shared low-rank axis and bounded by a norm constraint, which drastically reduces the number of free parameters and forces the model to use genomics only for targeted, stable adjustments of the class representations. This prevents the overfitting that plagues dense cross-modal interactions while still allowing each patient's molecular signature to refine the prototypical features.

Second, the **Class-Conditional Attention Aggregation** module. To connect global omics and local patches, we treat patch aggregation as an attention-based assignment problem. The calibrated prototypes serve as anchors that define a class-conditional prior for each slide. Individual patches are then assigned soft attention weights based on their similarity to the calibrated prototype of each candidate subtype, which amplifies subtype-relevant regions and suppresses irrelevant background, so that the genomic profile guides patch aggregation toward subtype-relevant regions for each class.

We validate OGPCMIL on three rare cancer cohorts from TCGA (UCS, THYM, SARC) under 1-shot, 5-shot, and 10-shot settings, using patch features extracted from three different pathology foundation models to ensure robustness. Our contributions are threefold. We formulate two under-explored challenges in few-shot multimodal MIL for rare cancers, namely cross-modal overfitting and global-local misalignment, and propose OGPCMIL as a unified, principle-driven solution. We introduce a low-rank prototype calibration scheme that tightly controls the capacity of genomic information flow, paired with a class-conditional attention aggregation that aligns global molecular profiles with localized morphological evidence. Extensive experiments demonstrate that OGPCMIL achieves state-of-the-art performance in 18 out of 27 comparisons across all three cohorts, and ablation studies confirm the critical role of both modules.[1]

## 2 Related Work

### 2.1 Pathology Foundation Models and MIL for Few-Shot WSI

Computational pathology has shifted from task-specific models to general-purpose representation learning. Recent pathology foundation models such as PLIP, CONCH, and KEEP (Huang et al., 2023; Lu et al., 2024; Zhou et al., 2026), which build on the vision-language pretraining pioneered by CLIP (Radford et al., 2021), provide transferable patch representations. Larger generalist models (Vorontsov et al., 2024; Wang et al., 2024; Ding et al., 2025) target common cancers and shift to rare domains, so we freeze them and adapt only a lightweight task-specific head. For gigapixel WSIs, multiple instance learning treats each slide as a bag of patches under slide-level weak super-

[1] Due to the page limit, architectural details and training configuration are provided in Appendix D, full baseline tables in Appendices A.1–A.3, hyper-parameter sensitivity in Appendix A.5, and the complete limitation discussion in Section 5.

vision. Attention-based aggregation (Ilse et al., 2018), patch-context modeling with Transformers (Shao et al., 2021; Vaswani et al., 2017), and CLAM-style attention (Lu et al., 2021) dominate the paradigm. Recent few-shot WSI methods introduce vision-language prompts (He et al., 2026) and multi-scale prompt tuning (Han et al., 2025), but these approaches rely mainly on visual morphology and yield class-agnostic global pooling that misses subtype-specific local nuances, a gap our class-conditional attention aggregation is designed to fill. Few-shot pathology classification has also been explored via prototype networks (Quan et al., 2024; Snell et al., 2017), meta-learning (Jiang et al., 2022), and vision-language MIL (Shi et al., 2024; Xu et al., 2025). Most of these methods fuse omics only superficially, whereas OGPCMIL calibrates prototypes with low-rank omics priors.

### 2.2 Multimodal Prototype Learning and Low-Rank Adaptation

Prototype learning matches samples to class prototypes, typically the mean of support features, circumventing large parametric classifiers and suiting few-shot settings. Injecting omics into prototypes follows two philosophies. Direct fusion concatenates or densely interacts high-dimensional omics with prototypes, introducing more degrees of freedom than few-shot data can constrain. Parameter-efficient adaptation injects cross-modal information with minimal trainable parameters, as exemplified by low-rank adaptation (LoRA), which constrains weight updates to a low-rank product (Hu et al., 2022). We adopt this principle by calibrating prototypes within a shared low-rank subspace. For multimodal fusion, optimal transport offers a rigorous framework for distribution matching under a cost matrix (Cuturi, 2013). We take this intuition to assign patch evidence to class prototypes via class-conditional attention, avoiding dense cross-modal interaction and its overfitting risk.

## 3 Method

In this section, we present OGPCMIL, our Omics-Guided Prototype Calibration Multimodal Multiple Instance Learning framework for few-shot rare cancer WSI subtyping. We first formulate the problem setting, then describe each component of the framework.

### 3.1 Problem Setting and Few-Shot Configuration

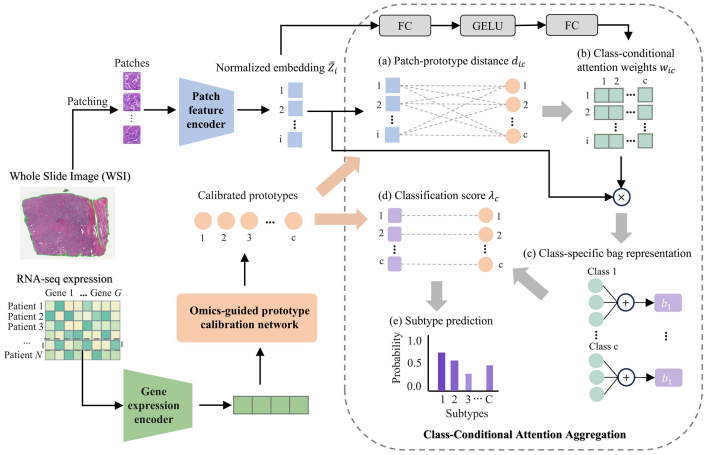



Figure 1: Overall architecture of OGPCMIL.

Figure 1 illustrates the overall architecture: a frozen encoder produces patch and omics representations, which the two modules below fuse into patient-specific prototypes and slide-level predictions.

The dataset is defined as $\mathcal{D} = \{(\mathcal{B}_i, y_i, \mathbf{g}_i)\}_{i=1}^{N_{\text{pat}}}$, where $\mathcal{B}_i = \{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,n_i}\}$ is the bag of patch features extracted from the $i$-th WSI by a frozen pre-trained pathology foundation model, $y_i \in \{1, \ldots, C\}$ denotes the slide-level subtype label, and $\mathbf{g}_i \in \mathbb{R}^G$ is the patient-level gene expression vector ($G = 19{,}938$ in our experiments).

In the $K$-shot setup, only $K$ labeled slides per class are available for training ($K \in \{1, 5, 10\}$). To prevent information leakage, all slides belonging to the same patient are placed in the same partition. Every setting is repeated 10 times, and we report the median as well as mean±standard deviation of balanced accuracy (BACC). This setup introduces two main difficulties. The first is *overfitting*, since directly fusing high-dimensional omics with thousands of patches has more degrees of freedom than $K$ samples can constrain. The second is *global-local misalignment*, since slide-level supervision never indicates where subtype-defining regions are, while omics captures a whole-tumor molecular summary.

### 3.2 Multimodal Feature Encoder

The encoder projects both modalities into a shared space using standard Transformer blocks (Vaswani et al., 2017). The patch branch applies a linear projection, LayerNorm, GELU, and dropout, then a single 4-head self-attention layer with a residual connection, and L2-normalizes the resulting evidence embeddings $\mathbf{z}_{i,j}$. The gene branch compresses the $G$-dimensional expression to $d_s = 64$ via two feed-forward layers with LayerNorm, GELU, dropout, and a Tanh-bounded state. Both modules are shown in Fig. 2 and 3. The pretrained foundation backbone is frozen and only these lightweight heads are trained.

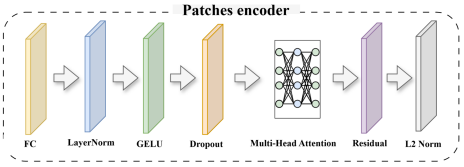



Figure 2: Patch feature encoding module.

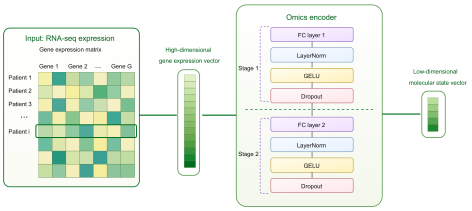



Figure 3: Gene expression encoding module.

### 3.3 Omics-Guided Low-Rank Prototype Calibration

**Prototype and semantic context construction.** Each class maintains a base prototype $\mathbf{p}_c \in \mathbb{R}^d$ (Snell et al., 2017) and $M = 16$ learnable context vectors $\mathbf{C}_c \in \mathbb{R}^{M \times d_c}$ ($d_c = 512$). As shown in Figure 4, the context vectors are flattened, passed through an MLP, added to the base prototype, and normalized to produce an uncalibrated prototype $\hat{\mathbf{p}}_c$.

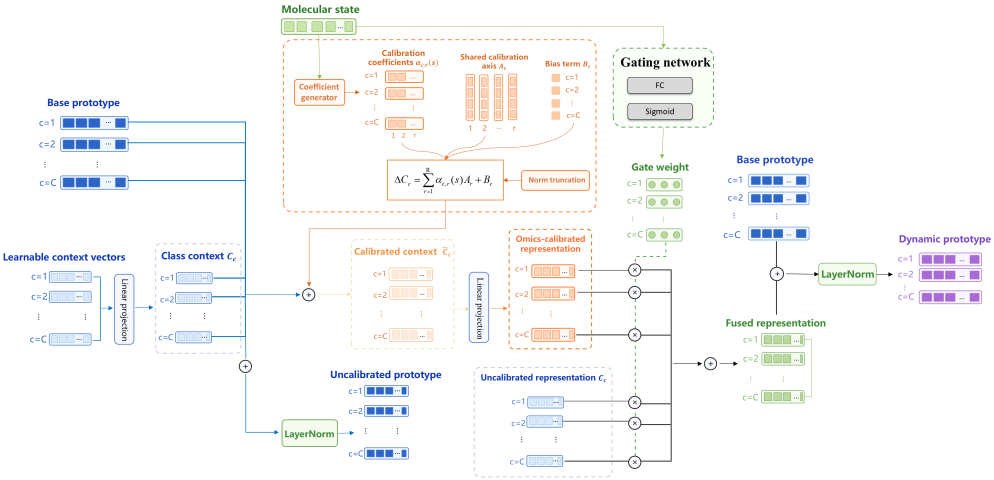



Figure 4: Omics-guided prototype calibration network.

**Low-rank calibration mechanism.** A two-layer MLP maps the molecular state $\mathbf{s}_i$ to per-class calibration coefficients $\alpha_{c,r}(\mathbf{s}_i)$ ($r = 1, \ldots, R$, with $R = 8$). The calibration offset for class $c$ and

patient $i$ is

$$\mathbf{\Delta}_{c,i} = \sum_{r=1}^{R} \alpha_{c,r}(\mathbf{s}_i)\mathbf{A}_r + \mathbf{B}_c \tag{1}$$

where $\mathbf{A}_r$ denotes the shared calibration axis, and $\mathbf{B}_c$ is the bias term. A norm truncation prevents excessive calibration:

$$\mathbf{\Delta}_{c,i} \leftarrow \mathbf{\Delta}_{c,i} \cdot \min\left(1, \frac{\delta}{\|\mathbf{\Delta}_{c,i}\|_F + \varepsilon}\right) \tag{2}$$

where $\delta = 2.0$ is the maximum allowed calibration norm, $\varepsilon$ prevents division by zero, and $\|\cdot\|_F$ denotes the Frobenius norm. The calibrated class context is

$$\tilde{\mathbf{C}}_c = \mathbf{C}_c + \mathbf{\Delta}_{c,i} \tag{3}$$

**Omics-gated fusion.** A gating network (linear layer + Sigmoid) outputs $\gamma_c \in [0, 1]$ for each class, blending uncalibrated and omics-calibrated representations:

$$\mathbf{u}_c^{\text{fuse}} = \gamma_c \tilde{\mathbf{u}}_c + (1 - \gamma_c)\mathbf{u}_c \tag{4}$$

where $\mathbf{u}_c$ is the uncalibrated class representation and $\tilde{\mathbf{u}}_c$ the omics-calibrated one. The fused representation is added to the base prototype followed by layer normalization, yielding the patient-specific dynamic prototype $\mathbf{p}_{i,c}$.

### 3.4 Class-Conditional Attention Aggregation

Rare cancer subtypes often differ only in a fraction of the tissue. Let $\bar{\mathbf{z}}_i$ denote the normalized embedding of the $i$-th patch and $\bar{\mathbf{p}}_c$ the normalized prototype of class $c$; their squared Euclidean distance is

$$d_{ic} = \|\bar{\mathbf{z}}_i - \bar{\mathbf{p}}_c\|_2^2 \tag{5}$$

A two-layer feed-forward network learns a class-agnostic importance bias $q_i$ per patch. The class-conditional attention weights are

$$w_{ic} = \frac{\exp(-d_{ic}/\tau + q_i)}{\sum_{j=1}^{N} \exp(-d_{jc}/\tau + q_j)} \tag{6}$$

with temperature $\tau = 0.2$. Each class obtains its own patch-weight distribution. The class-specific bag representation is

$$\mathbf{b}_c = \sum_{i=1}^{N} w_{ic}\bar{\mathbf{z}}_i \tag{7}$$

The classification score for class $c$ is determined by the match between the bag representation and the prototype:

$$\lambda_c = -\|\mathbf{b}_c - \bar{\mathbf{p}}_c\|_2^2 + f_m(\mathbf{b}_c) \tag{8}$$

where $f_m(\cdot)$ is a multilayer perceptron for fine-tuning the decision boundary. The final slide-level prediction is obtained by softmax over the scores.

### 3.5 Multi-constrained Joint Loss

Besides the calibrated branch, a base branch using only the uncalibrated prototypes acts as a training-time stability anchor. The total loss is a weighted sum of five terms:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + 0.25\mathcal{L}_{\text{base}} + 0.05\mathcal{L}_{\text{topo}} + 0.001\mathcal{L}_{\text{sparse}} + 0.1\mathcal{L}_{\text{cons}}. \tag{9}$$

where $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{base}}$ are cross-entropy losses, $\mathcal{L}_{\text{topo}}$ preserves prototype topology, $\mathcal{L}_{\text{sparse}}$ enforces low-rank sparsity on the calibration offset, and $\mathcal{L}_{\text{cons}}$ regularizes prediction consistency between the two branches. The topology and sparsity terms are

$$\mathcal{L}_{\text{topo}} = \frac{1}{C^2}\left\|D(\tilde{\mathbf{P}}) - D(\mathbf{P})\right\|_F^2 \tag{10}$$

$$\mathcal{L}_{\text{sparse}} = \frac{1}{d}\sum_{i=1}^{d}|\Delta_i| \tag{11}$$

with $D(\cdot)$ the $C \times C$ inter-class distance matrix and $\Delta_i$ the $i$-th element of the calibration offset $\mathbf{\Delta}_{c,i}$. Optimizer, learning rate, training epochs, and patch sampling are detailed in Appendix D.1.

# 4 EXPERIMENTS

We evaluate OGPCMIL on three rare cancer datasets under few-shot settings. We first describe the experimental setup, then present main comparison results against state-of-the-art baselines, conduct detailed ablation studies, and finally analyze generalization and robustness.

## 4.1 EXPERIMENTAL SETUP

We evaluate on three TCGA rare cancer datasets (Grossman et al., 2016). **UCS** has 2 subtypes with 52 WSIs from 31 patients (Cherniack et al., 2017), **THYM** has 4 subtypes with 153 WSIs from 94 patients (Radovich et al., 2018), and **SARC** has 5 subtypes with 575 WSIs from 240 patients (Network, 2017). The MPNST subtype of SARC has only 9 patients and is excluded from 10-shot. Gene expression profiles are RNA-seq from GDC. Patch tiling, gene normalization, and patient-level splits are detailed in Appendix D.2.

All slides of one patient stay in the same partition, with $K$ patients per class for training and the remainder as validation, and a fixed patient-level test set held constant across 10 repetitions. We report the BACC median with interquartile range, the BACC mean$\pm$std, the WF1 mean, and the AUC mean (Brodersen et al., 2010).

Patch features are extracted by three frozen foundation models. **PLIP** (Huang et al., 2023) produces 512-dimensional features, while **CONCH** (Lu et al., 2024) and **KEEP** (Zhou et al., 2026) both produce 768-dimensional features. The embedding dimension $d$ matches the foundation model output, and gene expression is compressed to $d_s = 64$ dimensions. We train with AdamW at a learning rate of $10^{-4}$ and weight decay of $10^{-4}$, batch size 2, for 20 epochs, sampling 2,048 patches per WSI per iteration.

We compare with five MIL methods, namely ABMIL (Ilse et al., 2018), TransMIL (Shao et al., 2021), MambaMIL (Yang et al., 2024), DGRMIL (Zhu et al., 2024), and PathPT (He et al., 2026). All use the same pre-extracted features and data splits, and baseline hyperparameters follow the original recommendations.

## 4.2 MAIN SOTA RESULTS ON RARE CANCERS

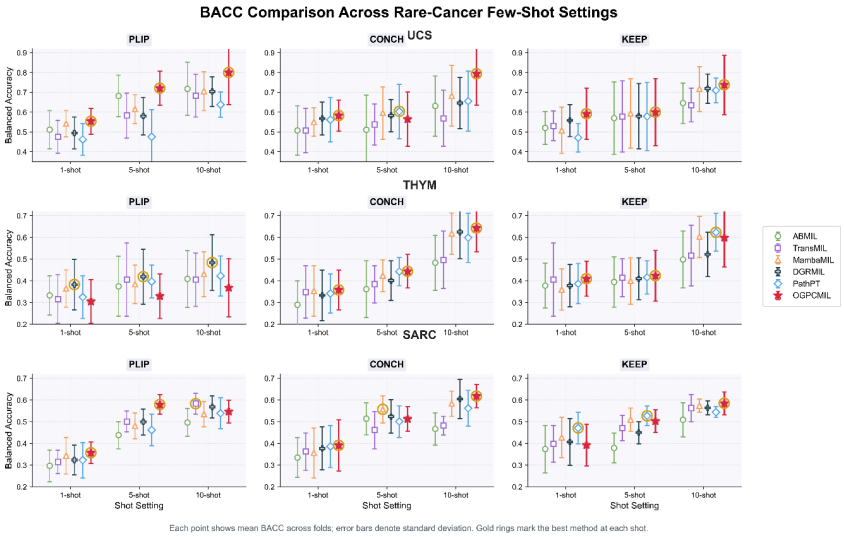



Figure 5: Performance comparison between OGPCMIL and mainstream MIL methods on three rare cancer datasets. Values are BACC mean; detailed mean$\pm$std results are reported in Appendices A.1–A.3.

Figure 5 presents the main comparison results across three datasets, three few-shot settings, and three foundation models, for 27 settings in total. OGPCMIL achieves the best BACC in 18 out of 27

settings, demonstrating consistent superiority across datasets and feature extractors. The complete numerical tables with mean±std for all settings are deferred to Appendices A.1–A.3.

Per dataset, OGPCMIL leads in 8/9 UCS settings, reaching 0.800 BACC with PLIP at 10-shot. It wins the hardest THYM cohort with CONCH at 0.643, and takes 5/9 SARC settings with the largest margin at PLIP 5-shot, 0.579 versus 0.501 for TransMIL. CONCH features give the most consistent results overall.

### 4.3 Integrated Ablation Experiments

We conduct detailed ablation studies on TCGA-UCS with CONCH features under the 10-shot setting, the configuration where OGPCMIL performs best.

#### 4.3.1 Prototype Calibration Network Ablation

Table 1 ablates the omics-guided prototype calibration network.

Table 1: Ablation of prototype calibration network on TCGA-UCS (CONCH, 10-shot).

| Setting | BACC mean±std | BACC med | WF1 | AUC |
|---|---|---|---|---|
| w/o omics | 0.750±0.101 | 0.750 | 0.682 | 0.766 |
| Late fusion | 0.697±0.136 | 0.707 | 0.651 | 0.708 |
| Full-rank calibration | 0.781±0.181 | 0.773 | 0.700 | 0.846 |
| w/o gate | 0.789±0.164 | 0.813 | 0.701 | 0.861 |
| **OGPCMIL** | **0.794±0.158** | **0.844** | **0.707** | **0.875** |

#### 4.3.2 Attention Aggregation Ablation

Table 2 compares the proposed class-conditional attention aggregation with conventional MIL aggregation strategies.

Table 2: Ablation of aggregation strategy on TCGA-UCS (CONCH, 10-shot).

| Aggregation | BACC mean±std | BACC med | WF1 | AUC |
|---|---|---|---|---|
| Mean pooling | 0.539±0.254 | 0.529 | 0.433 | 0.521 |
| Max pooling | 0.552±0.231 | 0.587 | 0.492 | 0.548 |
| Attention pooling | 0.604±0.217 | 0.630 | 0.525 | 0.625 |
| **OGPCMIL** | **0.794±0.158** | **0.844** | **0.707** | **0.875** |

#### 4.3.3 Loss Function Ablation

Table 3 ablates the three regularization terms.

Table 3: Ablation of loss components on TCGA-UCS (CONCH, 10-shot).

| Setting | BACC mean±std | BACC med | WF1 | AUC |
|---|---|---|---|---|
| w/o $\mathcal{L}_{\text{topo}}$ | 0.790±0.180 | 0.863 | 0.698 | 0.860 |
| w/o $\mathcal{L}_{\text{sparse}}$ | 0.787±0.165 | 0.823 | 0.694 | 0.858 |
| w/o $\mathcal{L}_{\text{cons}}$ | 0.781±0.172 | 0.837 | 0.690 | 0.849 |
| **OGPCMIL** | **0.794±0.158** | **0.844** | **0.707** | **0.875** |

Across the three ablations, each component proves necessary. Removing omics or using late fusion drops BACC to 0.750 and 0.697, and full-rank calibration raises variance to 0.181, confirming that low-rank omics calibration is key, as shown in Table 1. Class-conditional attention reaches 0.794 BACC, versus 0.539–0.604 for class-agnostic pooling, as shown in Table 2. All three regularizers aid stability, and removing $\mathcal{L}_{\text{cons}}$ causes the largest drop, to 0.781 BACC and 0.690 WF1, as reported in Table 3.

### 4.4 Hyper-parameter Sensitivity

We examine the sensitivity of the core hyper-parameters on the TCGA-SARC cohort with CONCH features. The low-rank dimension $R$ sets the number of shared calibration axes introduced in Eq. 1, controlling how much patient-specific genomic signal enters prototype refinement. The attention temperature $\tau$ controls the sharpness of the class-conditional attention weights in Eq. 6 and trades off targeted patch weighting against broad signal dispersion. The maximum calibration norm $\delta$ bounds the magnitude of prototype corrections in Eq. 2, preventing excessive distortion of the class representations. The prompt context length $M$ sets the capacity of the prototype context vectors.

We sweep $R \in \{1, 2, 4, 8, 12, 16\}$, $\tau \in \{0.05, 0.07, 0.1, 0.2, 0.4, 0.8\}$, $\delta \in \{0.5, 1.0, 2.0, 4.0\}$, and $M \in \{4, 8, 12, 16, 24, 32\}$. The default configuration uses $R$=8, $\tau$=0.2, $\delta$=2, and $M$=16, and the full numerical results are provided in Appendix A.5.

The prompt context length $M$ is the least sensitive hyper-parameter, with performance remaining within a narrow interval across all values. Rank $R$ is stable at 10-shot, with BACC between 0.605 and 0.618, but varies more at 1-shot, where extreme sample scarcity amplifies the effect of calibration capacity. The temperature $\tau$ and calibration norm $\delta$ are more sensitive, and performance degrades when they move far from the default values. The best settings coincide with the default hyper-parameters, confirming that the chosen configuration is robust.

### 4.5 Generalization and Robustness

#### 4.5.1 Multimodal Effectiveness

Figure 6 compares the full multimodal model with a unimodal variant that uses images only, across all 27 settings. In 23/27 settings, the multimodal model achieves higher BACC. The omics contribution is most consistent at 1-shot, where all 9 settings improve, indicating that genomic signals provide valuable complementary information when visual data is extremely limited.

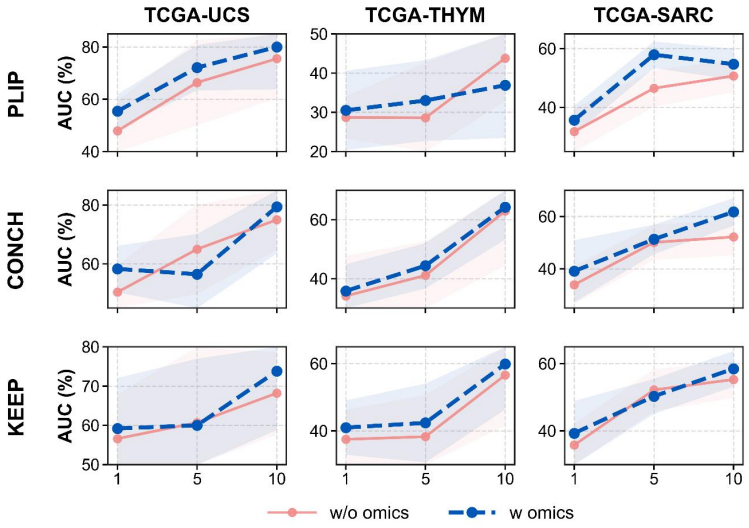



Figure 6: BACC comparison before and after removing the omics modality across three datasets, three foundation models, and three few-shot settings. Blue: multimodal (w/ omics); pink: unimodal (w/o omics).

The reliability diagrams in Fig. 10 of Appendix C.1 further show that, on TCGA-UCS with CONCH features at 10-shot, the vision-only variant is overconfident in high-confidence regions, whereas the full multimodal model tracks the diagonal more closely, demonstrating that omics guidance improves calibration.

#### 4.5.2 ZERO-SHOT BASELINE COMPARISON

Figure 11 in Appendix C.2 shows that zero-shot predictions from the three foundation models stay near random on these rare cancer tasks, while OGPCMIL improves substantially, reaching 10-shot BACC of 0.800, 0.794, and 0.738 on TCGA-UCS for PLIP, CONCH, and KEEP, respectively.

## 5 CONCLUSION

We presented OGPCMIL, a multimodal MIL framework for few-shot rare cancer subtyping that couples omics-guided low-rank prototype calibration with class-conditional attention aggregation. By treating genomics as a low-capacity calibrating prior and propagating it to patches through class-conditional attention, OGPCMIL mitigates cross-modal overfitting and global-local misalignment under extreme sample scarcity. On three rare cancer cohorts under 1-/5-/10-shot settings with three foundation models, it attains the best BACC in 18 of 27 comparisons, and ablations confirm the contribution of every component. A multimodal analysis pipeline further links subtyping to pathological region localization and pathway-level interpretation (Appendices B.1–B.2). We also evaluate OGPCMIL on a common cancer to verify that it does not degrade when training data is abundant; details are provided in Appendix A.6.

**Limitations and Future Work.** Our experiments rely exclusively on TCGA public data and lack multi-center external validation; generalization across institutions, scanners, and staining protocols remains to be assessed. Some rare subtypes involve very few samples, which may affect the stability of pathway interpretation and drug analysis, and region localization currently depends on internal evidence scores without systematic expert validation. Future work will incorporate multi-center cohorts, integrate spatial transcriptomics and proteomics, and study robust learning under missing modalities and lightweight clinical deployment.

### AI USE STATEMENT

In this work, we used generative AI tools for writing assistance, including drafting and polishing sections of the manuscript, and for code review, including verifying implementation details. We have not used generative AI tools for generating research ideas, conducting experiments, or producing experimental results. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

All whole-slide images and RNA-seq gene expression data used in this study were obtained from publicly available TCGA datasets. The use of these data follows the terms and conditions established by the dataset providers and adheres to applicable ethical and academic standards. No directly identifiable personal information is used in this study. We aim to advance computational pathology while respecting patient privacy and research integrity.

### REPRODUCIBILITY STATEMENT

To support reproducibility, we make our implementation, configuration files, and experimental settings publicly available at an anonymous repository: `https://anonymous.4open.science/r/OGPCMIL`. All datasets used in this study are publicly available, with data preparation instructions and patient-level splits provided in the repository. Our methodology and experimental settings are described in the main text, and further implementation details and full configurations are provided in the repository.

## REFERENCES

M. Ancona, E. Ceolini, C. Öztireli, and M. Gross. Towards better understanding of gradient-based attribution methods for deep neural networks, 2017. arXiv:1711.06104.

M. B. Binh, X. Sastre-Garau, L. Guillou, G. de Pinieux, P. Terrier, R. Lagacé, A. Aurias, I. Hostein, and J. M. Coindre. Mdm2 and cdk4 immunostainings are useful adjuncts in diagnosing well-differentiated and dedifferentiated liposarcoma subtypes: A comparative analysis of 559 soft tissue neoplasms with genetic data. *The American Journal of Surgical Pathology*, 29(10):1340–1347, 2005. doi: 10.1097/01.pas.0000170343.09562.39.

K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124, 2010.

R. J. Chen, M. Y. Lu, J. Wang, D. F. K. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2020. doi: 10.1109/TMI.2020.3021387.

R. J. Chen, M. Y. Lu, D. F. K. Williamson, T. Y. Chen, J. Lipkova, Z. Noor, M. Shaban, M. Shady, M. Williams, B. Joo, and F. Mahmood. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022. doi: 10.1016/j.ccell.2022.07.004.

A. D. Cherniack, H. Shen, V. Walter, C. Stewart, B. A. Murray, et al. Integrated molecular characterization of uterine carcinosarcoma. *Cancer Cell*, 31(3):411–423, 2017. doi: 10.1016/j.ccell.2017.02.010.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

C. E. DeSantis, J. L. Kramer, and A. Jemal. The burden of rare cancers in the united states. *CA: A Cancer Journal for Clinicians*, 67(4):261–272, 2017. doi: 10.3322/caac.21400.

M. A. Dickson, G. K. Schwartz, M. L. Keohan, S. P. D'Angelo, M. M. Gounder, P. Chi, C. R. Antonescu, J. Landa, L. X. Qin, A. M. Crago, S. Singer, A. Koff, and W. D. Tap. Progression-free survival among patients with well-differentiated or dedifferentiated liposarcoma treated with cdk4 inhibitor palbociclib: A phase 2 clinical trial. *JAMA Oncology*, 2(7):937–940, 2016. doi: 10.1001/jamaoncol.2016.0264.

T. Ding, S. J. Wagner, A. H. Song, R. J. Chen, M. Y. Lu, A. Zhang, A. J. Vaidya, G. Jaume, M. Shaban, A. Kim, D. F. K. Williamson, H. Robertson, B. Chen, C. Almagro-Pérez, P. Doucet, S. Sahai, C. Chen, C. S. Chen, D. Komura, A. Kawabe, M. Ochi, S. Sato, T. Yokose, Y. Miyagi, S. Ishikawa, G. Gerber, T. Peng, L. P. Le, and F. Mahmood. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31(11):3749–3761, 2025. doi: 10.1038/s41591-025-03982-3.

S. L. Freshour, S. Kiwala, K. C. Cotto, A. C. Coffman, J. F. McMichael, J. J. Song, M. Griffith, O. L. Griffith, and A. H. Wagner. Integration of the drug-gene interaction database (dgidb 4.0) with open crowdsource efforts. *Nucleic Acids Research*, 49(D1):D1144–D1151, 2021. doi: 10.1093/nar/gkaa1084.

M. Gillespie, B. Jassal, R. Stephan, M. Milacic, K. Rothfels, A. Senff-Ribeiro, J. Griss, C. Sevilla, L. Matthews, C. Gong, C. Deng, T. Varusai, E. Ragueneau, Y. Haider, B. May, V. Shamovsky, J. Weiser, T. Brunson, N. Sanati, L. Beckman, X. Shao, A. Fabregat, K. Sidiropoulos, J. Murillo, G. Viteri, J. Cook, S. Shorser, G. Bader, E. Demir, C. Sander, R. Haw, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 2022. doi: 10.1093/nar/gkab1028.

C. E. Gleason, M. A. Dickson, M. E. Klein Dooley, C. R. Antonescu, R. Gularte-Merida, M. Benitez, J. I. Delgado, R. P. Kataru, M. W. Y. Tan, M. Bradic, T. E. Adamson, K. Seier, A. L. Richards, M. Palafox, E. Chan, S. P. D'Angelo, M. M. Gounder, M. L. Keohan, C. M. Kelly, P. Chi, S. Movva, J. Landa, A. M. Crago, M. T. A. Donoghue, L. X. Qin, V. Serra, M. Turkekul, A. Barlas, D. M. Firester, K. Manova-Todorova, B. J. Mehara, M. Kovatcheva, N. S. Tan, S. Singer, W. D. Tap, and A. Koff. Therapy-induced senescence contributes to the efficacy of abemaciclib in patients with dedifferentiated liposarcoma. *Clinical Cancer Research*, 30(4):703–718, 2024. doi: 10.1158/1078-0432.CCR-23-2378.

R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12): 1109–1112, 2016. doi: 10.1056/NEJMp1607591.

M. Han, L. Qu, D. Yang, X. Zhang, X. Wang, and L. Zhang. Mscpt: Few-shot whole slide image classification with multi-scale and context-focused prompt tuning. *IEEE Transactions on Medical Imaging*, 44(9):3756–3769, 2025. doi: 10.1109/TMI.2025.3564976.

D. He, X. Zhou, W. Guan, L. Zhang, X. Zhang, S. Xu, G. Wang, L. Wang, X. Yuan, J. Ma, X. Sun, Y. Wang, K. Sun, Y. Zhang, and W. Xie. Boosting pathology foundation models via few-shot prompt-tuning for rare cancer subtyping. *Nature Communications*, 17(1):5058, 2026. doi: 10.1038/s41467-026-71715-2.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou. A visual-language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29(9):2307–2316, 2023. doi: 10.1038/s41591-023-02504-3.

M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pp. 2127–2136, 2018.

H. Jiang, M. Gao, H. Li, R. Jin, H. Miao, and J. Liu. Multi-learner based deep meta-learning for few-shot medical image classification. *IEEE Journal of Biomedical and Health Informatics*, 27 (1):17–28, 2022. doi: 10.1109/JBHI.2022.3215147.

M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, and M. Ishiguro-Watanabe. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(D1):D587–D592, 2023. doi: 10.1093/nar/gkac963.

J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795): 1929–1935, 2006. doi: 10.1126/science.1132939.

M. Lee. Recent advancements in deep learning using whole slide imaging for cancer prognosis. *Bioengineering*, 10(8):897, 2023.

A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo. The molecular signatures database (msigdb) hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015. doi: 10.1016/j.cels.2015.12.004.

M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. doi: 10.1038/s41551-020-00682-w.

M. Y. Lu, B. Chen, D. F. K. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, A. V. Parwani, A. Zhang, and F. Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. doi: 10.1038/s41591-024-02856-4.

M. Martens, A. Ammar, A. Riutta, A. Waagmeester, D. N. Slenter, K. Hanspers, R. A. Miller, D. Digles, E. N. Lopes, F. Ehrhart, L. J. Dupuis, L. A. Winckers, S. L. Coort, E. L. Willighagen, C. T. Evelo, A. R. Pico, and M. Kutmon. Wikipathways: connecting communities. *Nucleic Acids Research*, 49(D1):D613–D621, 2021. doi: 10.1093/nar/gkaa1024.

The Cancer Genome Atlas Research Network. Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell*, 171(4):950–965, 2017. doi: 10.1016/j.cell.2017.10.014.

M. Pizurica, Y. Zheng, F. Carrillo-Perez, H. Noor, W. Yao, C. Wohlfart, A. Vladimirova, K. Marchal, and O. Gevaert. Digital profiling of gene expression from histology images with linearized attention. *Nature Communications*, 15(1):9886, 2024. doi: 10.1038/s41467-024-54182-5.

H. Quan, X. Li, D. Hu, T. Nan, and X. Cui. Dual-channel prototype network for few-shot pathology image classification. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4132–4144, 2024. doi: 10.1109/JBHI.2024.3386197.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

M. Radovich, C. R. Pickering, I. Felau, G. Ha, H. Zhang, et al. The integrated genomic landscape of thymic epithelial tumors. *Cancer Cell*, 33(2):244–258, 2018. doi: 10.1016/j.ccell.2018.01.003.

B. Schmauch, A. Romagnoni, E. Pronier, C. Saillard, P. Maillé, J. Calderaro, A. Kamoun, M. Sefta, S. Toldo, M. Zaslavskiy, T. Clozel, M. Moarli, P. Courtiol, and G. Wainrib. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature Communications*, 11 (1):3877, 2020. doi: 10.1038/s41467-020-17678-4.

Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, and X. Ji. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

J. Shi, C. Li, T. Gong, Y. Zheng, and H. Fu. Vila-mil: Dual-scale vision-language multiple instance learning for whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11248–11258, 2024.

J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.

A. Vaidya, A. Zhang, G. Jaume, A. H. Song, T. Ding, S. J. Wagner, M. Y. Lu, P. Doucet, H. Robertson, C. Almagro-Perez, and R. J. Chen. Molecular-driven foundation model for oncologic pathology, 2025. arXiv:2501.16652.

W. T. A. van der Graaf, N. S. Heiss, C. L. Hynes, S. P. Keller, A. Weinman, J. Y. Blay, P. Franco, R. H. Giles, D. Lacombe, P. Schlatter, D. M. Thomas, S. B. van Waalwijk van Doorn-Khosrovani, M. Williamson, and R. Plummer. Overcoming the barriers to treatment of rare cancer patients in the era of precision oncology: A call to action. *Cancer Treatment Reviews*, 140:103013, 2025. doi: 10.1016/j.ctrv.2025.103013.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, K. Severson, E. Zimmermann, J. Hall, N. Tenenholtz, N. Fusi, E. Yang, P. Mathieu, A. van Eck, D. Lee, J. Viret, E. Robert, Y. K. Wang, J. D. Kunz, M. C. H. Lee, J. H. Bernhard, R. A. Godrich, G. Oakley, E. Millar, M. Hanna, H. Wen, J. A. Retamero, W. A. Moye, R. Yousfi, C. Kanan, D. S. Klimstra, B. Rothrock, S. Liu, and T. 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.

X. Wang, J. Zhao, E. Marostica, W. Yuan, J. Jin, J. Zhang, R. Li, H. Tang, K. Wang, Y. Li, F. Wang, Y. Peng, J. Zhu, J. Zhang, C. R. Jackson, J. Zhang, D. Dillon, N. U. Lin, L. Sholl, T. Denize, D. Meredith, K. L. Ligon, S. Signoretti, S. Ogino, J. A. Golden, M. P. Nasrallah, X. Han, S. Yang, and K. H. Yu. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, 2024. doi: 10.1038/s41586-024-07894-z.

Y. Xu, Y. Wang, F. Zhou, J. Ma, C. Jin, S. Yang, J. Li, Z. Zhang, C. Zhao, H. Zhou, Z. Li, H. Lin, X. Wang, J. Wang, A. Han, R. C. K. Chan, L. Liang, X. Zhang, and H. Chen. A multimodal knowledge-enhanced whole-slide pathology foundation model. *Nature Communications*, 16(1): 11406, 2025. doi: 10.1038/s41467-025-66220-x.

J. Xue and Q. Lyu. Challenges and opportunities in rare cancer research in china. *Science China Life Sciences*, 67(2):274–285, 2024. doi: 10.1007/s11427-023-2422-x.

S. Yang, Y. Wang, and H. Chen. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 296–306, 2024.

B. Zhang and H. Yue. Molecular subtyping of common solid tumors. *Cancer Research on Prevention and Treatment*, 43(11):1001–1004, 2016.

X. Zhou, L. Sun, D. He, W. Guan, G. Wang, R. Wang, L. Wang, X. Yuan, X. Sun, Y. Zhang, K. Sun, Y. Wang, and W. Xie. Knowledge-enhanced pretraining for vision-language pathology foundation model on cancer diagnosis. *Cancer Cell*, 44(4):777–791.e7, 2026. doi: 10.1016/j.ccell.2026.01.019.

W. Zhu, X. Chen, P. Qiu, A. Sotiras, A. Razi, and Y. Wang. Dgr-mil: Exploring diverse global representation in multiple instance learning for whole slide image classification. In *European Conference on Computer Vision*, pp. 333–351, 2024. doi: 10.1007/978-3-031-72920-1_19.

# A Full Experimental Tables

## A.1 UCS Dataset: Detailed Comparison

Table 4 presents the complete comparison results on the TCGA-UCS dataset across all methods, foundation models, and few-shot settings.

Table 4: TCGA-UCS: BACC mean±std for all methods. Best in **bold**, second-best underlined.

| Shot | Model | ABMIL | TransMIL | MambaMIL | DGRMIL | PathPT | **OGPCMIL** |
|---|---|---|---|---|---|---|---|
| 1-shot | PLIP | 0.512±0.096 | 0.476±0.083 | 0.542±0.066 | 0.495±0.081 | 0.462±0.080 | **0.554±0.065** |
| | CONCH | 0.508±0.125 | 0.508±0.111 | 0.551±0.071 | 0.568±0.083 | 0.562±0.113 | **0.583±0.079** |
| | KEEP | 0.521±0.083 | 0.531±0.076 | 0.508±0.117 | 0.558±0.081 | 0.471±0.073 | **0.592±0.129** |
| 5-shot | PLIP | 0.682±0.105 | 0.583±0.114 | 0.615±0.073 | 0.580±0.095 | 0.476±0.137 | **0.721±0.087** |
| | CONCH | 0.511±0.175 | 0.538±0.103 | 0.596±0.133 | 0.583±0.081 | 0.604±0.137 | 0.565±0.137 |
| | KEEP | 0.570±0.182 | 0.578±0.180 | 0.594±0.175 | 0.580±0.165 | 0.579±0.172 | **0.600±0.170** |
| 10-shot | PLIP | 0.718±0.133 | 0.683±0.107 | 0.706±0.098 | 0.704±0.074 | 0.638±0.065 | **0.800±0.162** |
| | CONCH | 0.631±0.152 | 0.569±0.141 | 0.682±0.153 | 0.646±0.130 | 0.656±0.151 | **0.794±0.158** |
| | KEEP | 0.646±0.102 | 0.636±0.085 | 0.717±0.113 | 0.719±0.074 | 0.710±0.062 | **0.738±0.150** |

## A.2 THYM Dataset: Detailed Comparison

Table 5 presents the complete comparison results on the TCGA-THYM dataset.

Table 5: TCGA-THYM: BACC mean±std for all methods. Best in **bold**, second-best underlined.

| Shot | Model | ABMIL | TransMIL | MambaMIL | DGRMIL | PathPT | **OGPCMIL** |
|---|---|---|---|---|---|---|---|
| 1-shot | PLIP | 0.333±0.090 | 0.316±0.112 | 0.365±0.086 | 0.383±0.117 | 0.325±0.098 | 0.305±0.101 |
| | CONCH | 0.290±0.110 | 0.349±0.121 | 0.353±0.116 | 0.333±0.117 | 0.342±0.091 | **0.358±0.091** |
| | KEEP | 0.378±0.103 | 0.406±0.169 | 0.360±0.095 | 0.378±0.098 | 0.387±0.093 | **0.410±0.081** |
| 5-shot | PLIP | 0.375±0.138 | 0.406±0.169 | 0.384±0.089 | 0.419±0.126 | 0.397±0.076 | 0.330±0.103 |
| | CONCH | 0.362±0.130 | 0.384±0.085 | 0.424±0.073 | 0.401±0.091 | 0.443±0.065 | **0.444±0.077** |
| | KEEP | 0.395±0.116 | 0.415±0.088 | 0.399±0.107 | 0.410±0.096 | 0.416±0.076 | **0.424±0.117** |
| 10-shot | PLIP | 0.409±0.131 | 0.405±0.123 | 0.431±0.104 | 0.484±0.129 | 0.423±0.092 | 0.369±0.134 |
| | CONCH | 0.483±0.127 | 0.497±0.132 | 0.618±0.095 | 0.625±0.122 | 0.599±0.113 | **0.643±0.110** |
| | KEEP | 0.498±0.131 | 0.517±0.140 | 0.603±0.094 | 0.522±0.103 | 0.623±0.087 | 0.599±0.135 |

## A.3 SARC Dataset: Detailed Comparison

Table 6 presents the complete comparison results on the TCGA-SARC dataset.

Table 6: TCGA-SARC: BACC mean±std for all methods. Best in **bold**, second-best underlined.

| Shot | Model | ABMIL | TransMIL | MambaMIL | DGRMIL | PathPT | **OGPCMIL** |
|---|---|---|---|---|---|---|---|
| 1-shot | PLIP | 0.297±0.073 | 0.315±0.053 | 0.344±0.084 | 0.324±0.069 | 0.323±0.081 | **0.357±0.049** |
| | CONCH | 0.335±0.092 | 0.363±0.086 | 0.356±0.114 | 0.377±0.099 | 0.386±0.097 | **0.391±0.119** |
| | KEEP | 0.374±0.110 | 0.398±0.085 | 0.427±0.094 | 0.407±0.108 | 0.471±0.073 | 0.393±0.096 |
| 5-shot | PLIP | 0.438±0.062 | 0.501±0.048 | 0.481±0.059 | 0.499±0.060 | 0.462±0.073 | **0.579±0.045** |
| | CONCH | 0.514±0.074 | 0.462±0.086 | 0.557±0.063 | 0.524±0.077 | 0.501±0.073 | 0.513±0.056 |
| | KEEP | 0.379±0.069 | 0.471±0.059 | 0.509±0.054 | 0.450±0.051 | 0.527±0.045 | 0.503±0.052 |
| 10-shot | PLIP | 0.496±0.064 | 0.583±0.049 | 0.535±0.058 | 0.568±0.051 | 0.539±0.071 | 0.547±0.053 |
| | CONCH | 0.467±0.074 | 0.482±0.043 | 0.583±0.058 | 0.605±0.090 | 0.562±0.082 | **0.618±0.053** |
| | KEEP | 0.509±0.079 | 0.563±0.062 | 0.574±0.031 | 0.564±0.031 | 0.544±0.025 | **0.585±0.053** |

### A.4 Omnibus Comparison Summary

Across all 27 comparisons (3 datasets × 3 shots × 3 foundation models), OGPCMIL achieves the best BACC in 18 settings, winning 8/9 UCS settings, 5/9 THYM settings, and 5/9 SARC settings.

### A.5 Hyper-parameter Sensitivity

This section studies the sensitivity of OGPCMIL to key hyper-parameters including rank $R$, temperature $\tau$, normalization scale $\delta$, and prompt context size $M$. We conduct experiments on TCGA-SARC dataset under multiple few-shot settings. Default settings are in bold. Metrics: BACC mean±std, median BACC (med), weighted F1 (WF1), AUC.

Table 7: Rank ($R$) sensitivity on TCGA-SARC (CONCH).

| Setting | $R$ | BACC mean±std | med | WF1 | AUC |
|---|---|---|---|---|---|
| 1-shot | 1 | 0.341±0.101 | 0.337 | 0.302 | 0.661 |
| | 2 | 0.349±0.093 | 0.337 | 0.335 | 0.663 |
| | 4 | 0.309±0.078 | 0.299 | 0.265 | 0.669 |
| | **8** | **0.391±0.119** | **0.385** | **0.346** | **0.683** |
| | 12 | 0.378±0.112 | 0.363 | 0.348 | 0.683 |
| | 16 | 0.327±0.063 | 0.318 | 0.291 | 0.653 |
| 5-shot | 1 | 0.493±0.078 | 0.516 | 0.522 | 0.760 |
| | 2 | 0.490±0.067 | 0.513 | 0.512 | 0.761 |
| | 4 | 0.480±0.085 | 0.496 | 0.505 | 0.759 |
| | **8** | **0.513±0.056** | **0.537** | **0.552** | **0.771** |
| | 12 | 0.478±0.051 | 0.522 | 0.508 | 0.758 |
| | 16 | 0.469±0.066 | 0.521 | 0.495 | 0.754 |
| 10-shot | 1 | 0.605±0.076 | 0.602 | 0.612 | 0.873 |
| | 2 | 0.611±0.061 | 0.618 | 0.627 | 0.890 |
| | 4 | 0.610±0.094 | 0.637 | 0.624 | 0.888 |
| | **8** | **0.618±0.094** | **0.622** | **0.640** | **0.893** |
| | 12 | 0.611±0.064 | 0.613 | 0.636 | 0.887 |
| | 16 | 0.609±0.102 | 0.625 | 0.627 | 0.880 |

Table 8: Temperature ($\tau$) sensitivity on TCGA-SARC (CONCH, 10-shot).

| $\tau$ | BACC mean±std | med | WF1 | AUC |
|---|---|---|---|---|
| 0.05 | 0.584±0.070 | 0.603 | 0.606 | 0.879 |
| 0.07 | 0.588±0.062 | 0.585 | 0.611 | 0.889 |
| 0.10 | 0.605±0.077 | 0.612 | 0.607 | 0.889 |
| **0.20** | **0.618±0.094** | **0.622** | **0.640** | **0.893** |
| 0.40 | 0.586±0.095 | 0.612 | 0.608 | 0.884 |
| 0.80 | 0.615±0.105 | 0.631 | 0.638 | 0.877 |

Table 9: Norm ($\delta$) sensitivity on TCGA-SARC (CONCH, 10-shot).

| $\delta$ | BACC mean±std | med | WF1 | AUC |
|---|---|---|---|---|
| 0.5 | 0.596±0.092 | 0.591 | 0.612 | 0.874 |
| 1.0 | 0.585±0.099 | 0.604 | 0.595 | 0.855 |
| **2.0** | **0.618±0.094** | **0.622** | **0.640** | **0.893** |
| 4.0 | 0.611±0.087 | 0.598 | 0.627 | 0.872 |

Table 10: Prompt context size ($M$) sensitivity on TCGA-SARC (CONCH, 10-shot).

| $M$ | BACC mean±std | med | WF1 | AUC |
|---|---|---|---|---|
| 4 | 0.607±0.101 | 0.629 | 0.625 | 0.891 |
| 8 | 0.613±0.049 | 0.615 | 0.621 | 0.897 |
| 12 | 0.615±0.079 | 0.625 | 0.633 | 0.892 |
| **16** | **0.618±0.094** | **0.622** | **0.640** | **0.893** |
| 24 | 0.614±0.069 | 0.606 | 0.638 | 0.882 |
| 32 | 0.618±0.101 | 0.630 | 0.636 | 0.894 |

### A.6 Applicability Boundary Analysis on Common Cancer

TCGA-BRCA was adopted for validation, owing to its large sample size within TCGA, and well-established histopathological distinctions between IDC and ILC subtypes reported in prior studies. This analysis is a verification that OGPCMIL does not degrade on common cancer when labels are plentiful, and that its performance attenuates under extreme few-shot constraints.

We use TCGA-BRCA for binary subtyping of invasive ductal carcinoma and invasive lobular carcinoma. Patch features are publicly available PLIP embeddings at 5x magnification from HiVE-MIL, whereas the main experiments extract 20x patches via TRIDENT. In both settings the backbone is frozen and outputs 512-dimensional features. Genomic features are RNA-seq expression profiles from GDC STAR-Counts. The full-shot regime uses 1,005 training slides with a single fold. The 10-shot regime uses 20 training slides, ten per class, across three independent folds. Apart from the magnification difference and the binary task configuration, the model architecture, training objective, and all hyper-parameters are kept identical to the main experiments.

Table 11: Applicability boundary analysis on TCGA-BRCA. Ten-shot results are mean ± std over three folds.

| Setting | ACC | BACC | WF1 | AUC |
|---|---|---|---|---|
| Full-shot | 0.9930 | 0.9920 | 0.9931 | 0.9994 |
| 10-shot | 0.659±0.084 | 0.703±0.074 | 0.693±0.076 | 0.773±0.071 |

Full-shot training is evaluated on a single fold and achieves near-saturated performance without modality-induced degradation. The 10-shot regime is repeated over three independent folds to ensure statistical reliability under higher variance. On the IDC versus ILC task, performance drops substantially under 10-shot constraints, indicating that extreme sample scarcity poses inherent challenges for subtyping even when morphological hallmarks are well documented. This observation aligns with the performance gains from omics-guided calibration in the ablation studies, jointly supporting the value of omics guidance in scenarios with limited visual supervision.

## B Supplementary Visualizations

This appendix presents the interpretability analysis of OGPCMIL, following the protocol of the original study, which analyzes the TCGA-SARC cohort with CONCH pathology features under the 10-shot setting, using the test-set patients. Specifically, we (i) localize critical pathological regions from patch-level evidence scores, and (ii) attribute model predictions to genes and aggregate the attribution at the pathway level.

### B.1 Pathological Region Localization

To avoid the contingency of a single slide, we export patch-level evidence scores for all test slides. For each whole slide image, the model first obtains the predicted subtype, and the patch contributions with respect to that predicted class serve as the explanation target; the evidence scores of all patches are normalized to $[0, 1]$. A higher normalized score indicates stronger support of the patch for the current prediction. For cross-slide comparison, we report the per-slide mean score, the top-5% threshold, the top-1% threshold, and the enrichment fold of high-evidence regions. Two correctly predicted slides are selected as representatives for each predicted subtype of SARC.

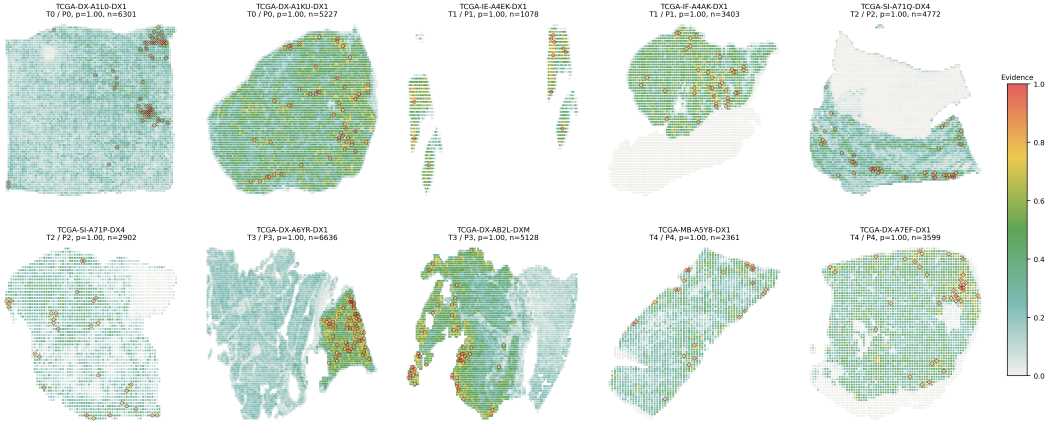



Figure 7: Pathological region localization heatmaps on representative test slides. Colors denote normalized evidence scores, and red hollow circles mark the top-1% patches within each slide.

Figure 7 shows the localization results for 10 representative test slides covering five subtypes. High-evidence patches form locally enriched regions instead of being uniformly scattered, showing that the model focuses on a small number of high-contribution regions. Quantitatively, the top-5% enrichment fold of the representative slides is mostly above 2, so a few local high-evidence regions dominate the prediction.

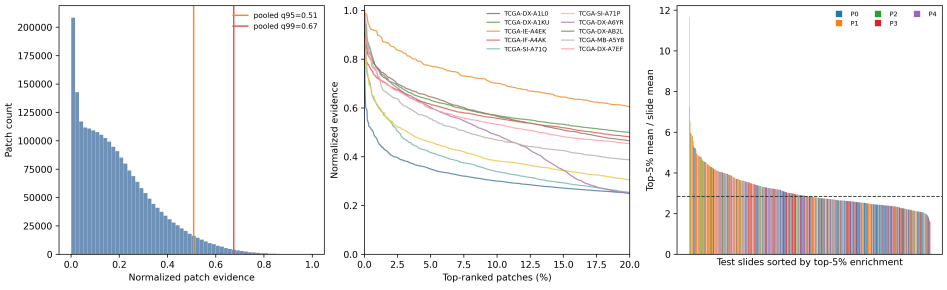



Figure 8: Evidence score distribution and enrichment across the test cohort. Left: distribution of patch-level evidence scores over all test slides; middle: per-slide scores sorted in descending order; right: top-5% enrichment fold of each test slide.

Figure 8 further validates this sparsity at the cohort level. The test set contains 2,368,020 valid patches in total, with an overall mean normalized evidence score of 0.195, an overall top-5% threshold of 0.508, an overall top-1% threshold of 0.675, and a median top-5% enrichment fold of 2.844. Most patches concentrate in the low-score interval, and only a small fraction reaches the high-evidence thresholds. The sorted score curve drops rapidly within the first few percentiles. These results confirm that the model is able to focus on critical regions from a large number of candidate patches.

### B.2 Gene Attribution and Pathway-Level Interpretation

For molecular interpretation, we apply gradient×input attribution (Ancona et al., 2017) to the patient-level gene expression vector, measuring the contribution of each gene to the current pre-

diction by its absolute attribution value. For each patient, gene attribution scores are mapped onto pathway sets (MSigDB Hallmark (Liberzon et al., 2015), Reactome (Gillespie et al., 2022), KEGG (Kanehisa et al., 2023), and WikiPathways (Martens et al., 2021)), and the mean attribution of the matched genes within a pathway is taken as the pathway contribution score, retaining the number of matched genes and the main contributing genes. High-contribution pathways are then aggregated per predicted subtype, and their occurrence frequency within the subtype is computed to identify stable pathways with subtype consistency. In addition, the abnormal direction of each pathway is estimated by comparing the mean expression of its genes between patients of the same predicted subtype and other patients, marking pathways as relatively activated or suppressed.

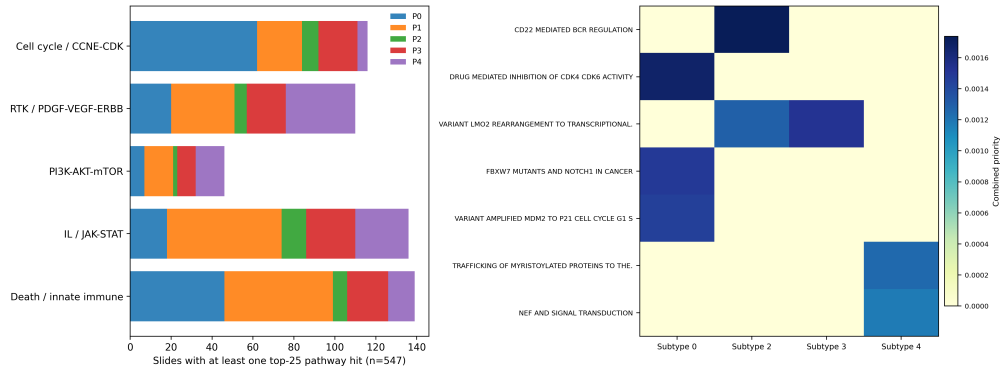



Figure 9: Gene attribution aggregated at the pathway and mechanism levels. Left: occurrence counts of mechanism modules in case-level high-score pathways; right: priority heatmap of strictly filtered subtype-stable pathways (darker indicates higher priority). Subtypes 0–4 correspond to DDLPS, LMS, MPNST, MFS, and UPS, respectively.

Figure 9 aggregates the attribution results at the mechanism and pathway levels. At the mechanism level, cell cycle/CCNE-CDK (62 occurrences), death/innate immunity (46), and RTK/PDGF-VEGF-ERBB (20) dominate in predicted subtype 0; IL/JAK-STAT and death/innate immunity stand out in subtype 1; RTK/PDGF-VEGF-ERBB (34) and IL/JAK-STAT (26) are prominent in subtype 4. At the pathway level, the most salient stable pathways of subtype 0 (DDLPS) include CDK4/CDK6 activity-related pathways and the MDM2-p21 cell-cycle pathway. The CDK4/CDK6 regulation-related pathway is activated in 36 DDLPS-predicted patients (frequency 0.450), consistent with the well-documented CDK4/MDM2 amplification in dedifferentiated liposarcoma (Binh et al., 2005; Dickson et al., 2016; Gleason et al., 2024). We further cross-check the top pathways against drug-gene interaction databases (DGIdb (Freshour et al., 2021), CMap (Lamb et al., 2006)) to connect subtype mechanisms with candidate targeted agents. These results show that the model's molecular attribution aggregates into pathway-level mechanisms with reproducible frequency and direction, providing traceable interpretive evidence for the model's predictions.

# C Supplementary Results

This appendix provides two supplementary result figures referenced in Section 4.5.

## C.1 Reliability Diagrams

Figure 10 (left) shows the reliability diagrams referenced in Section 4.5.

## C.2 Zero-shot Comparison

Figure 11 (right) compares zero-shot predictions from the three foundation models with OGPCMIL.

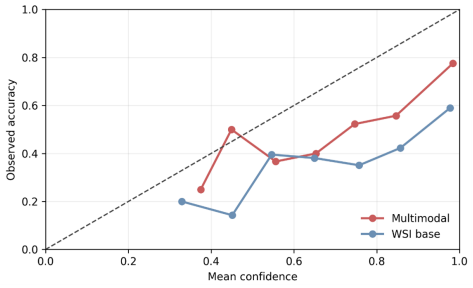



Figure 10: Reliability diagrams before and after removing the omics modality on TCGA-UCS (CONCH, 10-shot). Blue: vision-only; red: full OGPCMIL. Dashed diagonal: perfect calibration.

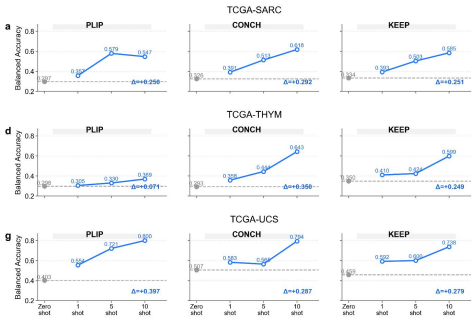



Figure 11: Zero-shot subtyping BACC of the three foundation models compared with OGPCMIL under 1-, 5-, and 10-shot settings on three rare cancer datasets. On TCGA-UCS at 10-shot, OGPCMIL exceeds zero-shot by +0.297 (PLIP), +0.287 (CONCH), and +0.219 (KEEP).

# D Implementation Details

## D.1 Training Details

All trainable parameters are optimized end-to-end with AdamW (learning rate $1 \times 10^{-4}$, weight decay $1 \times 10^{-4}$), batch size 2, for 20 epochs, sampling 2,048 patches per WSI per iteration; the checkpoint with the highest validation BACC is selected for evaluation. We fix the random seed to 42 and adopt patient-level data splits as described in Section 4.1. The five loss terms use weights $\lambda_{\text{cls}} = 1.0$, $\lambda_{\text{base}} = 0.25$, $\lambda_{\text{topo}} = 0.05$, $\lambda_{\text{sparse}} = 0.001$, and $\lambda_{\text{cons}} = 0.1$.

## D.2 Data Preprocessing and Splits

Patch tiling: WSIs are tiled into $224 \times 224$ patches at $20\times$ magnification via TRIDENT (Vaidya et al., 2025). Gene expression (RNA-seq from GDC, $G = 19{,}938$ genes) is log-transformed and z-scored. Patient-level splitting: all WSIs of one patient are placed in the same partition; for each $K$-shot setting a fixed patient-level test set is held constant across 10 repetitions, $K$ patients per class are sampled for training, and remaining patients form the validation set.