One Slide, Many Views: Unifying Complementary Foundation Model Perspectives for WSI Analysis

Yihui Wang[1], Yingxue Xu[1], Shu Yang[1], Yequan Bie[1], Jiabo Ma[1], Fengtao Zhou[1], and Hao Chen[1,2,3,4,5⋆]

[1] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong SAR, China
[2] Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[3] State Key Laboratory of Molecular Neuroscience, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[4] Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[5] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China
{ywangrm,yxueb,syangcw,ybie,jmabq,fzhouaf}@connect.ust.hk, jhc@ust.hk

Abstract. Recent advances in foundation models (FMs) have led to a proliferation of Whole Slide Image (WSI) feature extractors, yet WSI analysis still faces a practical challenge: no single FM consistently performs best across all tasks, and the performance gap among different Multiple Instance Learning (MIL) aggregators built on a single strong FM can become limited in some settings. Recent research further shows that simple ensembles of FMs can outperform the best individual FM, indicating that different FMs capture partially overlapping yet complementary information. This motivates a shift from selecting a single FM toward coordinating and exploiting multiple FMs jointly. However, existing multi-FM approaches mostly rely on naive feature concatenation or patch-level selection, operating on raw correlated FM features without explicitly disentangling shared and view-specific information. To address this challenge, we reformulate FM ensembling as a Multi-View Learning (MVL) problem. We propose Multi-View Multiple Instance Learning (MV-MIL), a plug-and-play framework that treats features from each FM as a distinct view of a WSI. MV-MIL introduces an information-theoretic disentanglement module to exploit cross-view complementarity while suppressing redundancy, together with a Multi-Branch Prediction Head that integrates shared and view-specific representations within MIL pipelines. Extensive experiments on 14 public benchmarks show that MV-MIL consistently achieves state-of-the-art performance, with its plug-and-play design yielding robust gains across diverse MIL methods and FM combinations.

Keywords: Computational Pathology · Multiple Instance Learning · Foundation Models

⋆ Corresponding author.

1 Introduction

In recent years, deep learning has revolutionized the field of Computational Pathology (CPath) [4, 18, 23]. The analysis of Whole Slide Images (WSIs) is central to CPath, driving automation in cancer diagnosis, prognosis prediction, and treatment response assessment [7, 17]. Due to their gigapixel size, WSIs are typically processed using a Multiple Instance Learning (MIL) framework [16, 21, 25, 33, 36, 46, 47], where the slide is first divided into thousands of patches, each independently encoded into a low-dimensional embedding using a pretrained feature extractor. These patch-level embeddings are then aggregated to make slide-level predictions. More recently, the field of computational pathology has been rapidly advanced by the emergence of large-scale foundation models (FMs) [5, 15, 24, 27, 31, 32, 38, 40, 44, 45, 49] that learn rich representations from histopathology images, serving as powerful feature extractors for MIL pipelines.

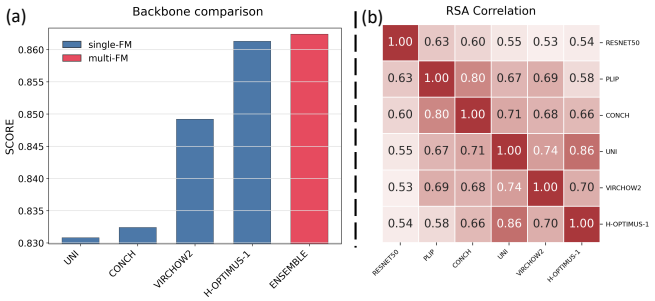

Fig. 1: (a) A simple average ensemble can outperform all single FMs on the CRC-Molecular task. (b) Representational Similarity Analysis of the FMs. The varied correlation coefficients confirm that the models learn diverse and non-redundant representations.

However, the rapid proliferation of FMs presents a new challenge for downstream applications: **how should we choose and utilize these models**? Different FMs are trained with diverse data, network architectures, and pretraining strategies. Consequently, they capture distinct biological signals and exhibit different representational biases [12]. For a given downstream task, no single FM can guarantee optimal performance across datasets [30]. For instance, a model like H-Optimus-1 may demonstrate strong performance on breast cancer tasks, but its efficacy can be less competitive on liver cancer-related tasks [28]. Recent studies further show that the robustness of FMs substantially narrows the performance gap between different MIL architectures built on top of a single FM [22, 29]. While recent benchmarks [28, 30] show that even simple ensembles of FMs yield consistent gains, such naive combinations may not fully unlock their potential. To better understand the feature interactions among them, we

perform a representational similarity analysis [19] across FMs (Fig. 1; details in the *Supplementary Material*), which reveals heterogeneous pairwise correlations and empirically shows that different FMs learn partially shared yet view-specific representations, giving rise to both redundancy and complementarity. Taken together, these observations suggest that **instead of designing more complex MIL heads on a single FM, it is more promising to investigate how multiple FMs can be coordinated and exploited jointly.** Previous work has explored simple fusion strategies, such as averaging, feature concatenation [30] or attention-based ensemble [11], which yield modest performance gains over individual models. Recent approaches like AdaFusion [42] introduce adaptive mechanisms to select the most relevant FM-derived feature at the patch level. However, these approaches share a fundamental limitation: they operate directly on raw, highly correlated FM features and implicitly treat all information as equally informative. As a result, they fail to distinguish redundant signal shared across FMs from truly complementary cues, allowing dominant shared components to overshadow more subtle view-specific information. Thus, the core challenge is not merely to ensemble multiple FMs, but to separate their shared semantic content from FM-specific signals and then recombine them in a task-aware way. This structure is precisely the setting studied in **Multi-View Learning** (MVL), where multiple correlated views of the same object are assumed to contain both common and view-specific information.

Motivated by this perspective, we reformulate FM ensembling as a multi-view multiple instance learning problem and propose *Multi-View Multiple Instance Learning* (MV-MIL), a plug-and-play framework that explicitly ensembles multiple FMs within a unified multi-view MIL formulation. For each patch, the features extracted from different FMs are regarded as distinct **views** that capture both overlapping and complementary information. To better exploit cross-view complementarities while suppressing redundancy, we model these views with an *information-theoretic disentanglement* module inspired by the Information Bottleneck (IB) principle [37]. This module decomposes them into a shared component and a set of view-specific components, yielding predictive yet non-redundant representations by explicitly balancing sufficiency for the downstream task with compression. The resulting shared and specific features are then fed into a standard MIL pipeline. To fully exploit both components, we introduce a *Multi-Branch Prediction Head*, in which each branch is instantiated as a MIL aggregator operating on either the shared or the view-specific features. The branch outputs are then fused at the slide level, and the underlying MIL blocks can be instantiated by any existing architecture (e.g., ABMIL, DSMIL, CLAM), making the design plug-and-play across different MIL backbones. Our main contributions are summarized as follows:

- **Reformulating FM ensemble as a multi-view learning problem.** We formulate the ensemble of multiple FMs as a structured Multi-View Learning task, where each FM is treated as a distinct semantic view of the same WSI.
- **Introducing an information bottleneck framework for better ensembling of FMs.** We propose a view-wise IB formulation to decompose

FM outputs into shared and view-specific components, enabling compact, predictive, and non-redundant representations.
- **Designing a plug-and-play module compatible with various MIL pipelines.** Our MV-MIL framework, including the Multi-Branch Prediction Head, can be seamlessly integrated with different MIL methods and FMs.
- **Achieving strong performance across diverse pathology tasks.** Extensive experiments on molecular prediction and survival analysis demonstrate state-of-the-art results across multiple public benchmarks.

2 Related Work

2.1 Foundation Models in CPath

The field of CPath has been fundamentally transformed by the advent of FMs, which leverage vast amounts of data to improve diagnostic accuracy and prognostic assessment. CTransPath [40] is a pioneering model, followed by a proliferation of models, such as UNI [5], which leverages DINOv2 for self-supervised learning, and GPFM [27], which explores knowledge distillation to enhance generalizability. In parallel, vision-language FMs have also emerged, including PLIP [15] and CONCH [24], which align pathology images with text to enable promptable and zero-shot transfer. Such FMs can encode semantic information that differs from purely vision-pretrained models, providing another source of representation diversity across FMs. Multimodal FMs such as mSTAR [44] further integrate WSI-derived features with molecular and clinical context for pretraining. More recently, industrial efforts have pushed the boundaries of model size, producing large-scale models like the Virchow series [38, 49] and the H-Optimus series [31,32]. The rapid proliferation of models, each with distinct training data, modalities, pretraining objectives, and inductive biases, has created a new challenge: no single model consistently outperforms the others [28,30]. This motivates our work to move beyond selecting one model and instead develop a principled framework for integrating the complementary insights from multiple FMs.

2.2 Multiple Instance Learning

Given the gigapixel size of WSIs, the predominant approach for WSI analysis is to first generate patch-level features via a feature extractor and then employ a MIL model to aggregate these features for prediction. Previously, CPath research often relied on ResNet50 [14] for feature extraction and designed complex MIL architectures to squeeze out performance gains. However, the advent of powerful FMs has changed this dynamic. In practice, high-quality FM features shrink the relative importance of the MIL model: simple ABMIL [16] often matches or surpasses more elaborate designs. Recent studies have therefore explored complementary ways to improve the representations consumed by MIL models. DGR-MIL [48] enhances bag-level representation by learning diverse global representations within WSIs. R^2T [36] re-embeds instance features to

restore fine-grained cues and consistently improves MIL performance, closing the gap between ResNet50 features and FM-level features. Cracking Instance Jigsaw Puzzles [6] further explores inter-instance semantic correlations beyond the standard permutation-invariant MIL formulation. Beyond these MIL-side advances, recent studies have also investigated how to compose features from multiple FMs. COBRA [20] utilizes a novel contrastive strategy to pretrain a slide-level encoder directly from multiple FMs' features, generating a unified representation that notably remains compatible with unseen FMs at inference time. AdaFusion [42] further follows this trend by performing prompt-guided fusion of multiple FMs, delivering consistent gains over single-model baselines. Inspired by meta-analysis, the Meta-Encoder [11] also integrates features from different FMs using self-attention strategies to achieve gains, particularly on complex gene expression prediction tasks. ELF [26] further provides a systematic ensemble learning framework for combining pathology FMs in precision oncology, showing that multi-FM integration can improve downstream prediction.

2.3 Multi-view Learning with Information Theory

Features produced by multiple FMs can be viewed as complementary **views** of the same slide. Classical multi-view learning seeks to capture cross-view consensus while preserving view-specific signals, with correlation-based objectives such as CCA and its deep and variational forms encouraging aligned subspaces across views [2, 39]. Information-theoretic approaches, notably the information bottleneck and its deep variants, make this trade-off explicit by maximizing mutual information with labels to ensure sufficiency while compressing nuisance variation through limits on input dependence [1,35,37]. Recent multi-view IB formulations [34, 41] further separate shared components that are consistent across views from private components that encode complementary evidence, often with explicit regularizers for cross-view redundancy. Our work follows this line: we treat the outputs of several FMs as multiple views, use an IB-guided objective to decompose them into a common subspace and view-specific subspaces before aggregation, and keep the downstream MIL head simple.

3 Methods

3.1 Problem Formulation

WSI analysis is naturally framed as a MIL problem. In this paradigm, a WSI is treated as a bag $\mathcal{B}$, which contains a set of N patches $\{p_i\}_{i=1}^N$. A pretrained FM f maps each patch to a D-dimensional feature vector $x_i = f(p_i) \in \mathbb{R}^D$, and we collect the instance set as $X = \{x_i\}_{i=1}^N$. Following standard MIL approaches [16], the goal is to learn a model $\mathcal{M}$ that maps the instance set X to the bag label $\hat{Y}$. This process can be formally defined as:

$$\hat{Y} \leftarrow \mathcal{M}(X) := \mathcal{C}(\mathcal{A}(X)) = \mathcal{C}(\mathcal{A}(\{x_i\}_{i=1}^N))$$

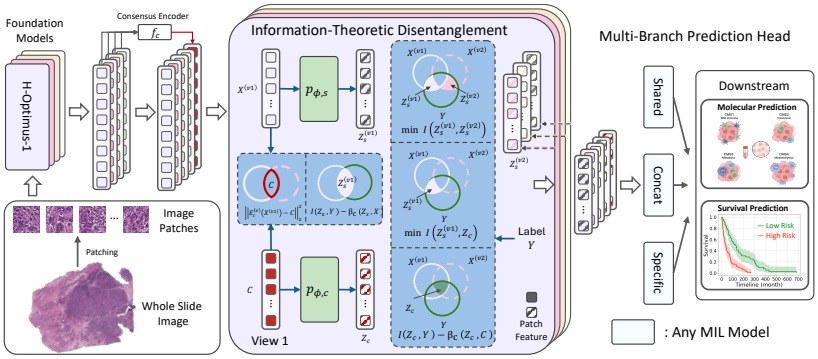

Fig. 2: Overview of our proposed multi-view WSI analysis framework. We first extract patch-level features from V different FMs, resulting in V distinct views. These views are then processed by our **Disentanglement Module**, which decomposes them into a shared representation Z_c and V view-specific representations $\{Z_s^{(v)}\}_{v=1}^{V}$. The disentangled features are subsequently fed into a **Multi-Branch Prediction Head**, which consists of multiple independent MIL branches that aggregate and classify the shared, specific, and concatenated representations. The final bag-level prediction is obtained by averaging the logits from all branches.

where $\mathcal{A}(\cdot)$ is typically a function that aggregates the set of N instance features X into a single bag-level representation and $\mathcal{C}(\cdot)$ is a classifier that maps the bag representation to the final prediction $\hat{Y}$.

In our work, we extend this framework to a **Multi-View** setting. We treat the features extracted from V different FMs for the same WSI as V distinct views. Formally, for a WSI bag $\mathcal{B}$, we have V sets of patch-level features $\{X_v\}_{v=1}^{V}$, where $X_v = \{x_i^{(v)}\}_{i=1}^{N} \in \mathbb{R}^{N \times d_v}$ represents the N patch features extracted by the v-th FM. Our goal is to learn a multi-view MIL model $\mathcal{M}_{\text{multi}}$ that fuses these views to predict the WSI-level label Y.

$$\hat{Y} \leftarrow \mathcal{M}_{\text{multi}}(X_1, \ldots, X_V) := \mathcal{C}(\mathcal{A}(\mathcal{F}(X_1, \ldots, X_V)))$$

In this multi-view formulation, the MIL aggregator $\mathcal{A}(\cdot)$ and bag classifier $\mathcal{C}(\cdot)$ follow the same definitions as in standard MIL. The role of this function $\mathcal{F}(\cdot)$ is to map the V input views to a single, unified patch-level representation. Unlike traditional methods that directly concatenate features from all views, we model $\mathcal{F}$ as an **Information-Theoretic Disentanglement** process. This process first decomposes all views into a shared component $Z_c = \{z_{c,i}\}_{i=1}^{N}$ and V specific components $Z_s^{(v)} = \{z_{s,i}^{(v)}\}_{i=1}^{N}$. The subsequent sections detail the design of the information-theoretic objectives required to learn Z_c and $\{Z_s^{(v)}\}_{v=1}^{V}$ effectively.

3.2 Information-Theoretic Disentanglement

As illustrated in Fig. 2, we first use V FMs to extract patch-level features and obtain $X_v = \{x_i^{(v)}\}_{i=1}^N$ for $v = 1, \ldots, V$. In our implementation, we adopt four widely used pathology FMs as view encoders, namely UNI [5], CONCH [24], H-Optimus-1 [32], and Virchow2 [49]. To fuse these multi-view features, we propose the **Information-Theoretic Disentanglement Module** $\mathcal{F}(\cdot)$ that decomposes the input features into shared and specific components. Before feeding the multi-view features into the disentanglement module, we apply a stochastic view dropout strategy during training to improve robustness to missing or corrupted views. Given V views with instance-level features $\{X_v\}_{v=1}^V$, where $X_v \in \mathbb{R}^{N \times d_v}$, we independently sample a binary gate for each view and construct the masked inputs as

$$\tilde{X}_v = (1 - m_v)\, X_v, \qquad m_v \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p_{\text{drop}}), \quad v = 1, \ldots, V.$$

For each view v, $\mathcal{F}$ applies parallel shared and specific encoders to map $\tilde{X}_v$ into a per-patch shared projection $\tilde{Z}_c^{(v)}$ and a view-specific representation $Z_s^{(v)}$. To jointly learn a shared component Z_c and the set $\{Z_s^{(v)}\}_{v=1}^V$, we optimize two complementary objectives: a Consistency Objective $\mathcal{L}_{\text{consistency}}$ that enforces cross-view agreement among $\{\tilde{Z}_c^{(v)}\}_{v=1}^V$ and aligns them to Z_c, and a Specificity Objective $\mathcal{L}_{\text{specific}}$ that encourages each $Z_s^{(v)}$ to be predictive for its own view while suppressing cross-view leakage.

Consistency Objective In designing the learning objective for the common representation, our primary requirement is to recover information that is truly shared across V views, rather than averaging pairwise overlaps as the final representation. Theoretically, the purest form of shared information is the Gács-Körner (GK) common information [10]. The GK common information is defined as the largest component that can be deterministically recovered from all views simultaneously. However, the GK common information is notoriously difficult to compute and is task-agnostic. We therefore learn a tractable, task-relevant approximation via a consensus encoder f_c. Concretely, for each patch i in a WSI bag $\mathcal{B}$, we concatenate the aligned multi-view features $\{x_i^{(v)}\}_{v=1}^V$ and compute a consensus token

$$c_i = f_c\Big(\text{Concat}\big(x_i^{(1)}, \ldots, x_i^{(V)}\big)\Big), \quad c_i \in \mathbb{R}^{d_c}.$$

Stacking all N patches yields the consensus representation $C \in \mathbb{R}^{N \times d_c}$, which serves as the concrete implementation of the shared component used by our framework. The learning of this C is guided by the GK common information principle of finding a component that is both high in entropy and consistent across all views. Specifically, we have V encoders $\{E_c^{(v)}\}_{v=1}^V$, each dedicated to extracting shared representations from its respective view. We then optimize the following objective to fulfill the GK principle:

$$\max_{C} \; H(C) - \sum_{v=1}^{V} \mathbb{E}[\|E_c^{(v)}(\tilde{X}_v) - C\|_2^2]$$

where the first term $H(C)$ is the Shannon entropy of C, encouraging C to be maximally informative, while the second term ensures that C is recoverable from each view's shared encoder output.

After obtaining the consensus representation C, we infer a per-patch shared latent representation via a variational bottleneck, $z_{c,i} \sim p_\phi(z_{c,i} \mid c_i)$, yielding $Z_c = \{z_{c,i}\}_{i=1}^N$. In WSI analysis, supervision is provided only at the slide level, while a slide contains a large and variable number of patches with substantial tissue heterogeneity. To produce a scalable slide-level representation from Z_c under this weakly supervised setting, we introduce a WSI-oriented aggregation operator $\mathcal{A}_c(\cdot)$ and define $\bar{Z}_c = \mathcal{A}_c(\{z_{c,i}\}_{i=1}^N)$. This leads to the following IB objective:

$$\max_{\phi} \; I(\bar{Z}_c; Y) \; - \; \beta_c \, I(Z_c; C), \qquad z_{c,i} \sim p_\phi(z_{c,i} \mid c_i),$$

where $I(A; B)$ describes the mutual information between variables A and B. In practice, both terms in this IB objective are intractable to compute directly. We therefore optimize tractable bounds and approximations for each term, which are implemented as follows:

$$\begin{aligned} \mathcal{L}_{\text{IB-Consistency}} = \; & \underbrace{\mathbb{E}_{(C,Y)} \, \mathbb{E}_{p_\phi(Z_c|C)} \big[-\log q_\psi\big(Y \mid \bar{Z}_c\big) \big]}_{\text{sufficiency of } \bar{Z}_c \text{ for } Y} \\ & + \; \underbrace{\beta_c \, \mathbb{E}_{p(C)} \, \mathrm{KL}\big(p_\phi(Z_c \mid C) \,\|\, r_\eta(Z_c)\big)}_{\text{compression of } Z_c \text{ w.r.t. } C}, \end{aligned}$$

The first term maximizes a variational lower bound on $I(\bar{Z}_c; Y)$ via the Barber–Agakov bound [3], encouraging $\bar{Z}_c$ to be sufficient for predicting Y. The second term is a VIB-style regularizer [1] that upper-bounds $I(Z_c; C)$ by replacing the intractable marginal with a tractable prior $r_\eta(Z_c)$, which compresses Z_c with respect to C. Expectations are estimated by averaging over patches within each WSI bag. Detailed derivations are provided in the *Supplementary Material*. By combining the practical implementations of the GK objective and the IB objective, our final $\mathcal{L}_{\text{consistency}}$ is formulated as a single loss to be minimized:

$$\begin{aligned} \mathcal{L}_{\text{consistency}} = & \underbrace{\sum_{v=1}^{V} \mathbb{E}[\|E_c^{(v)}(\tilde{X}_v) - C\|_2^2] - H(C)}_{\text{GK Objective}} \\ & \underbrace{-I(\bar{Z}_c; Y) + \beta_c I(Z_c; C)}_{\text{IB Objective}} \end{aligned}$$

Specificity Objective Each view provided by an FM may contain specific information not present in other views. If the model focuses only on the shared representation while ignoring this specific information, it may fail to fully leverage the complementary strengths of each view [9]. We therefore introduce a view-specific representation $Z_s^{(v)}$ for each view v to capture information that is unique to that view and not already in the common representation. By modeling these view-specific features alongside the shared representation, the overall multi-view model can leverage complementary information from each view, leading to a more comprehensive and discriminative representation for downstream tasks. In summary, the specificity module aims to preserve predictive but view-specific signals, while ensuring that $Z_s^{(v)}$ does not overlap with the shared features Z_c. For each view v, we aggregate per-patch specific features into a slide-level representation via $\bar{Z}_s^{(v)} = \mathcal{A}_s^{(v)}\big(\{z_{s,i}^{(v)}\}_{i=1}^N\big)$. Similar to the consistency objective, we formalize the specificity objective via an IB objective:

$$\max_{\phi} \sum_{v=1}^{V} (I(\bar{Z}_s^{(v)}; Y) - \beta_s I(Z_s^{(v)}; X_v)) \quad z_{s,i}^{(v)} \sim p_\phi(z_{s,i}^{(v)} \mid x_i^{(v)})$$

Both terms in this objective are intractable to compute directly. We therefore optimize a tractable variational bound by approximating each term:

$$\mathcal{L}_{\text{IB-Specific}} = \sum_{v=1}^{V} \Big(\underbrace{\mathbb{E}_{(X_v,Y)} \mathbb{E}_{p_\phi(Z_s^{(v)}|X_v)} \big[-\log q_{\psi_s^{(v)}}(Y \mid \bar{Z}_s^{(v)}) \big]}_{\text{sufficient for } Y \text{ (from view } v)} + \underbrace{\beta_s \, \mathbb{E}_{p(X_v)} \mathrm{KL}\Big(p_\phi(Z_s^{(v)} \mid X_v) \,\|\, r_{\eta_s^{(v)}}(Z_s^{(v)}) \Big)}_{\text{compress w.r.t. } X_v} \Big)$$

We instantiate both $r_\eta(Z_c)$ and $r_{\eta_s^{(v)}}(Z_s^{(v)})$ as factorized Gaussian variational priors, following standard VIB practice; this yields tractable KL upper bounds while encouraging compact latent representations without imposing task-specific prior assumptions. Meanwhile, to enforce disentanglement between the shared and view-specific components and to ensure different views learn distinct specific signals, we add the following penalty:

$$\mathcal{L}_{\text{orthogonal}} = \sum_{u \neq v} I(Z_s^{(u)}; Z_s^{(v)}), \qquad \mathcal{L}_{\text{disentangle}} = \sum_{v=1}^{V} I(Z_c; Z_s^{(v)})$$

The orthogonality term $\mathcal{L}_{\text{orthogonal}}$ encourages different view-specific representations to be as independent as possible by minimizing the mutual information between $Z_s^{(u)}$ and $Z_s^{(v)}$ for $u \neq v$. The disentanglement term $\mathcal{L}_{\text{disentangle}}$ further reduces redundancy between the shared representation Z_c and each view-specific representation $Z_s^{(v)}$, promoting view-specific components that capture information complementary to Z_c. Since estimating mutual information is intractable

in practice, we replace $I(\cdot;\cdot)$ with a covariance-based surrogate that suppresses linear dependence:

$$\mathcal{L}_{\text{cov}}(A, B) \;=\; \left\| \frac{1}{N-1} (A - \bar{A})^\top (B - \bar{B}) \right\|_F^2$$

Combining these components, our final specificity loss to be minimized is:

$$\mathcal{L}_{\text{specific}} = \underbrace{\mathcal{L}_{\text{IB-Specific}}}_{\text{IB Objective}} + \underbrace{\lambda_1 \sum_{u \neq v} \mathcal{L}_{\text{cov}}(Z_s^{(u)}, Z_s^{(v)})}_{\text{Orthogonality Penalty}} + \underbrace{\lambda_2 \sum_{v=1}^{V} \mathcal{L}_{\text{cov}}(Z_c, Z_s^{(v)})}_{\text{Disentanglement Penalty}}$$

3.3 Aggregation and Classification

As detailed in Sec. 3.2, our disentanglement module $\mathcal{F}(\cdot)$ produces two distinct sets of patch-level representations: the shared component Z_c and the V specific components $\{Z_s^{(v)}\}_{v=1}^{V}$. Our model is designed to process these components using a flexible **Multi-Branch Prediction Head**. Instead of a single aggregation step, this head creates multiple, independent MIL branches. Each branch consists of its own aggregation function $\mathcal{A}(\cdot)$ and classification function $\mathcal{C}(\cdot)$. Specifically (shown in Fig. 2), we instantiate $V + 2$ independent MIL branches: Shared Branch: One branch that takes only the shared component Z_c as input:

$$\hat{Y}_{\text{shared}} = \mathcal{C}_{\text{sh}}(\mathcal{A}_{\text{sh}}(Z_c))$$

Specific Branches: V branches, where each branch takes only its corresponding specific component $Z_s^{(v)}$ as input:

$$\hat{Y}_{\text{spec}}^{(v)} = \mathcal{C}_{\text{sp}}^{(v)}(\mathcal{A}_{\text{sp}}^{(v)}(Z_s^{(v)}))$$

Concat Branch: One branch that takes the full concatenated representation $Z_{\text{fused}} = \text{Concat}(Z_c, Z_s^{(1)}, \ldots, Z_s^{(V)})$ as input:

$$\hat{Y}_{\text{concat}} = \mathcal{C}_{\text{cat}}(\mathcal{A}_{\text{cat}}(Z_{\text{fused}}))$$

The final bag-level prediction $\hat{Y}$ is the unweighted average of the logits from all $V + 2$ branches:

$$\hat{Y} = \frac{1}{V+2} \left(\hat{Y}_{\text{shared}} + \sum_{v=1}^{V} \hat{Y}_{\text{spec}}^{(v)} + \hat{Y}_{\text{concat}} \right)$$

The shared branch uses the cross-FM consensus Z_c for a robust prediction, the view-specific branches preserve cues unique to each FM, and the concat branch captures their joint interactions. Averaging their logits thus acts as a

structured ensemble of distinct decision pathways, avoiding both shared-only over-compression and naive-fusion redundancy (Table 3). Given the ensemble logits $\hat{Y}$, we adopt the **Cross-Entropy** loss for the classification task and the **Negative Log-Likelihood Survival** loss [8] for the survival prediction task. Thus, the final training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \alpha_c \mathcal{L}_{\text{consistency}} + \alpha_s \mathcal{L}_{\text{specific}}$$

where α_c and α_s balance the contributions of each loss component.

4 Experiments

4.1 Datasets and Experimental Settings

Datasets: We evaluate our method across two key tasks: molecular prediction and survival analysis. (1) **Molecular Prediction**. We evaluate our method on eight public datasets. From the BCNB [43] cohort, we perform three binary receptor status prediction tasks: BCNB-ER, BCNB-PR, and BCNB-HER2. We also use five public TCGA subsets: BRCA-Molecular for PAM50 subtyping; BRCA-TP53 for TP53 mutation prediction; CRC-Molecular for consensus molecular subtype classification; CRC-KRAS for KRAS mutation prediction; and SKCM-BRAF for BRAF mutation prediction. (2) **Survival Analysis**. We use six TCGA cohorts (BRCA, CRC, KIRC, LUAD, LUSC, and SKCM) to predict patient survival outcomes from WSIs. Further details on all cohorts are provided in the *Supplementary Material*.

Evaluation Metrics: For molecular prediction, we primarily report the Area Under the Curve (AUC) due to the significant class imbalance often present in WSI cohorts. For survival prediction, we use the standard Concordance index (C-index) [13] to evaluate prognostic performance. All experiments are conducted using a 5-fold cross-validation methodology, and we report the mean and standard deviation of the results across all folds to ensure a robust evaluation. We provide runtime and memory comparisons in the *Supplementary Material*.

Compared Methods: We first establish a set of single FM baselines. These consist of several different feature extractors, with the resulting patch embeddings uniformly aggregated by a standard ABMIL [16] head for all models. The extractors include: (1) a classic ImageNet-pretrained ResNet50 [14] as a conventional baseline, and (2) five FMs: PLIP [15], UNI [5], CONCH [24], Virchow2 [49], and H-Optimus-1 [32]. Furthermore, we compare against three multi-FM baselines, which serve as our primary competitors. (1) Naive Concat: This method first concatenates the features from all views at the patch-level, creating a single, high-dimensional feature vector. This unified set of patch features is then fed into a standard ABMIL. (2) Self-Attn [11]: This baseline first applies the same concatenation as Naive Concat. It then processes these fused patch features through a standard self-attention layer and feeds the resulting features into the final ABMIL aggregator. (3) Naive Ensemble: This method averages the final

prediction logits from each of the independent single-view ABMIL models defined in the previous section. *To ensure fairness and reproducibility, we restrict our comparisons to methods with public code, thereby excluding AdaFusion [42].*

Foundation Model Selection: All methods, including ours, utilize features from four strong FMs (UNI, CONCH, Virchow2, and H-Optimus-1), chosen for their robust and widely validated performance in recent benchmarks [28, 30].

Implementation Details: WSI preprocessing follows CLAM [25]. All experiments are conducted on a single NVIDIA H800 GPU. Hyperparameter settings are provided in the *Supplementary Material*. Code is publicly available at `https://github.com/wyhsleep/MV-MIL`.

4.2 Main Results

Molecular Prediction. As shown in Table 1, our method attains the highest AUC on all eight datasets, with an average of **0.8008**. On average, we outperform the strongest multi-view baseline (Naive Ensemble, 0.7811) by +**1.97** points, and exceed Naive Concat and Self-Attn by +**2.92** and +**3.29** points, respectively. Compared with the best single-view foundation model on average (H-Optimus-1, 0.7821), our approach improves by +**1.87** points. The gains are consistent across receptor status (ER, PR, HER2), subtyping (PAM50, CRC-CMS), and mutation targets (TP53, KRAS, BRAF).

Survival Analysis. Our method achieves the highest average C-Index of **0.6905**, outperforming the strongest multi-view baseline (Naive Concat, 0.6748) by +**1.57** points. It also exceeds Naive Ensemble (0.6686) and Self-Attn (0.6588) by +**2.19** and +**3.17** points, and improves over the best single-view foundation model on average (Virchow2, 0.6691) by +**2.14** points.

Table 1: Main results for Molecular Prediction and Survival Analysis. Our model is compared against individual FMs and baseline multi-view methods. **Bold** indicates the best performance, and underline indicates the second-best.

Datasets	ResNet50 [14]	PLIP [15]	UNI [5]	CONCH [24]	Virchow2 [49]	H-Optimus-1 [32]	Naive Concat	Naive Ensemble	Self-Attn [11]	Ours
	+ABMIL									
Molecular Prediction (AUC)										
BCNB-ER	0.8008±0.0326	0.8381±0.0455	0.9040±0.0101	0.8738±0.0334	0.9125±0.0052	0.9259±0.0239	0.9242±0.0162	0.9237±0.0231	0.9154±0.0074	**0.9322±0.0136**
BCNB-HER2	0.6782±0.0557	0.7344±0.0201	0.7424±0.0246	0.7556±0.0345	0.7397±0.0215	0.7683±0.0277	0.7597±0.0243	0.7655±0.0258	0.7643±0.0108	**0.7894±0.0192**
BCNB-PR	0.7266±0.0373	0.7808±0.0410	0.8377±0.0282	0.8170±0.0270	0.8464±0.0347	0.8628±0.0261	0.8454±0.0396	0.8562±0.0380	0.8448±0.0301	**0.8770±0.0287**
BRCA-Molecular	0.6605±0.0587	0.7311±0.0307	0.7611±0.0132	0.7794±0.0242	0.7656±0.0251	0.7759±0.0240	0.7681±0.0266	0.7747±0.0085	0.7708±0.0252	**0.7905±0.0285**
BRCA-TP53	0.6825±0.0435	0.7839±0.0532	0.8213±0.0468	0.8263±0.0513	0.8283±0.0477	0.8451±0.0345	0.8527±0.0352	0.8442±0.0388	0.8337±0.0307	**0.8654±0.0229**
CRC-KRAS	0.5629±0.0543	0.5656±0.0688	0.6094±0.0803	0.6243±0.0633	0.5512±0.0607	0.5945±0.0904	0.5542±0.0857	0.6247±0.0974	0.5428±0.0962	**0.6450±0.0761**
CRC-Molecular	0.5894±0.0604	0.6929±0.0554	0.8308±0.0328	0.8324±0.0258	0.8492±0.0350	0.8613±0.0282	0.8602±0.0292	0.8624±0.0353	0.8573±0.0308	**0.8685±0.0272**
SKCM-BRAF	0.5556±0.0360	0.5482±0.0269	0.6072±0.0563	0.5928±0.0508	0.6255±0.0723	0.6232±0.0858	0.6084±0.0432	0.5972±0.0511	0.6138±0.0513	**0.6385±0.0641**
Average (Mol)	0.6571	0.7094	0.7642	0.7627	0.7648	0.7821	0.7716	0.7811	0.7679	**0.8008**
Survival Analysis (C-index)										
TCGA-BRCA	0.6044±0.0396	0.6134±0.0613	0.6705±0.0230	0.6844±0.0342	0.7093±0.0530	0.7006±0.0293	0.7017±0.0432	0.7009±0.0343	0.6978±0.0409	**0.7098±0.0354**
TCGA-CRC	0.5807±0.0482	0.6719±0.0128	0.7110±0.0264	0.7292±0.0315	0.7126±0.0303	0.6982±0.0243	**0.7321±0.0179**	0.7221±0.0269	0.6844±0.0370	**0.7321±0.0219**
TCGA-KIRC	0.6204±0.0926	0.7060±0.0821	0.7051±0.0505	0.7363±0.0412	0.7169±0.0689	0.7124±0.0558	0.7343±0.0701	0.7203±0.0581	0.7340±0.0602	**0.7445±0.0146**
TCGA-LUAD	0.6126±0.0202	0.6236±0.0392	0.6251±0.0396	0.6411±0.0432	0.6554±0.0295	0.6595±0.0468	0.6605±0.0514	0.6390±0.0486	0.6261±0.0319	**0.6693±0.0462**
TCGA-LUSC	0.5909±0.0798	0.5578±0.0488	0.6116±0.0562	0.5987±0.0658	0.5907±0.0476	0.6076±0.0492	0.5983±0.0416	0.6124±0.0481	0.6027±0.0668	**0.6311±0.0587**
TCGA-SKCM	0.5952±0.0388	0.6011±0.0483	0.6340±0.0325	0.6116±0.0241	0.6296±0.0238	0.6159±0.0115	0.6219±0.0343	0.6167±0.0220	0.6075±0.0180	**0.6561±0.0686**
Average (Surv)	0.6007	0.6290	0.6596	0.6669	0.6691	0.6657	0.6748	0.6686	0.6588	**0.6905**

4.3 Ablation Studies

Plug-and-play generality across MIL methods. To assess whether our framework is tied to a particular MIL architecture, we plug the disentanglement module into five standard MIL methods: ABMIL, DSMIL, MeanMIL, MaxMIL, and CLAM. Fig. 3a summarizes the average molecular-prediction AUC when using the best single-FM baseline for each backbone (Best FM) versus our multi-view variant (Ours). The improvements are particularly pronounced for DSMIL and MaxMIL, indicating that our multi-view modeling and information-theoretic disentanglement can reliably enhance diverse MIL designs without modifying their core structure.

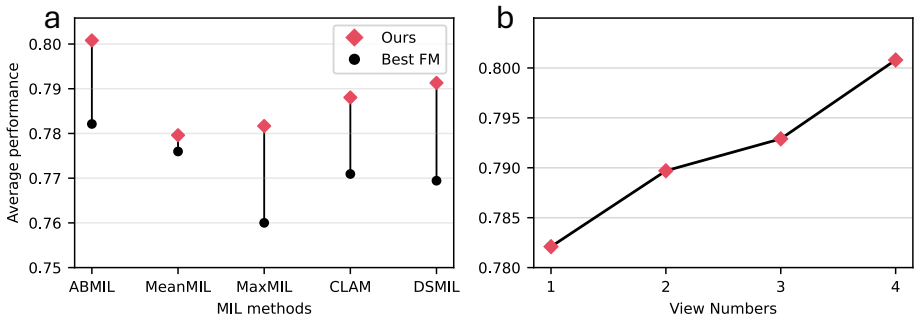

Fig. 3: **(a)** Across different MIL backbones, our multi-view framework consistently improves over the best single FM. **(b)** Performance increases as more FM views are used.

Plug-and-play generality across FM backbones. We conduct a leave-one-out ablation study to investigate the individual contribution of each FM (view) to our final model's performance. As shown in Table 2, our full model achieves the highest average AUC of 0.8008, while all four leave-one-out variants (0.7779–0.7930) underperform the full model. This indicates that every view provides a non-redundant contribution and that our method effectively integrates information from all FMs. Among the four models, removing H-Optimus-1 causes the largest performance degradation (a drop of 0.0229, from 0.8008 to 0.7779), which is consistent with its role as the strongest single-view baseline (0.7821). Nevertheless, the full multi-FM model still outperforms H-Optimus-1 alone, as well as all leave-one-out variants, showing that simply selecting the strongest FM is suboptimal and that explicitly modeling and exploiting complementary information across multiple FMs is crucial for achieving the best performance. Furthermore, as illustrated in Fig. 3b, we observe a clear monotonic increase in the best achievable AUC as the number of integrated views grows. Specifically, the performance scales steadily from 0.7821 with a single FM (H-Optimus-1) to 0.7897 when combining two models (CONCH and H-Optimus-1), and further

improves to 0.7929 with three models (CONCH, Virchow2, and H-Optimus-1). Ultimately, integrating all four FMs yields the highest AUC of 0.8008. This continuous improvement directly validates our method's ability to extract additive value from each newly introduced FM. In addition, the competitive performance of all leave-one-out configurations indicates that our framework can robustly adapt to different FM subsets, supporting its use as a general plug-and-play module for arbitrary FM combinations.

Table 2: Ablation study on the choice of foundation models for molecular prediction (AUC). Columns labeled "w/o ..." correspond to leave-one-out variants of our multi-view model, where one FM is removed while the remaining three are used. "Ours" denotes the full model using all four FMs.

Datasets	UNI [5]	CONCH [24]	Virchow2 [49]	H-Optimus-1 [32]	w/o UNI	w/o CONCH	w/o Virchow2	w/o H-Optimus-1	Ours
BCNB-ER	0.9040±0.0101	0.8738±0.0334	0.9125±0.0052	0.9259±0.0239	**0.9326±0.0161**	0.9230±0.0178	0.9305±0.0114	0.9114±0.0183	0.9322±0.0136
BCNB-HER2	0.7424±0.0246	0.7556±0.0345	0.7397±0.0215	0.7683±0.0277	0.7738±0.0257	0.7760±0.0219	0.7724±0.0163	0.7667±0.0217	**0.7894±0.0192**
BCNB-PR	0.8377±0.0282	0.8170±0.0270	0.8464±0.0347	0.8628±0.0261	0.8727±0.0311	0.8716±0.0322	0.8613±0.0287	0.8570±0.0366	**0.8770±0.0287**
BRCA-Molecular	0.7611±0.0132	0.7794±0.0242	0.7656±0.0251	0.7759±0.0240	**0.8051±0.0266**	0.7812±0.0343	0.7708±0.0234	0.7650±0.0285	0.7905±0.0285
BRCA-TP53	0.8213±0.0468	0.8263±0.0513	0.8283±0.0477	0.8451±0.0345	0.8501±0.0414	0.8444±0.0466	0.8469±0.0401	0.8516±0.0431	**0.8654±0.0229**
CRC-KRAS	0.6094±0.0803	0.6243±0.0633	0.5512±0.0607	0.5945±0.0904	0.6204±0.0983	0.6410±0.0874	0.6341±0.0696	0.6283±0.1155	**0.6450±0.0761**
CRC-Molecular	0.8308±0.0328	0.8324±0.0258	0.8492±0.0350	0.8613±0.0282	0.8655±0.0238	0.8632±0.0235	0.8618±0.0310	0.8370±0.0398	**0.8685±0.0272**
SKCM-BRAF	0.6072±0.0563	0.5928±0.0508	0.6255±0.0723	0.6232±0.0858	0.6234±0.1013	0.6280±0.0649	0.6152±0.0588	0.6064±0.0348	**0.6385±0.0641**
Average (Mol)	0.7642	0.7627	0.7648	0.7821	0.7930	0.7911	0.7866	0.7779	**0.8008**

Effect of prediction branches. The left panel of Table 3 evaluates how different prediction branches affect molecular prediction. Using only the view-specific branch already yields a strong average AUC of 0.7834, which is better than the best single-FM baseline (H-Optimus-1, 0.7821), indicating that most discriminative signals indeed reside in the FM-specific components. Adding the concatenated representation of shared and specific features (Concat + Specific) brings a consistent improvement to 0.7900 on average, showing that the joint space of shared and specific information is beneficial. Finally, the full head that aggregates shared, concatenated, and specific branches achieves the best overall performance with an average AUC of 0.8008. These results confirm that our multi-branch design successfully leverages both shared information and view-specific signals, rather than relying on either component alone.

Effect of consistency and specificity objectives. The right panel of Table 3 analyzes the role of the two information-theoretic losses used in our disentanglement module. Removing the consistency term (w/o $\mathcal{L}_{\text{consistency}}$) degrades the average molecular AUC from 0.8008 to 0.7902, while removing the specificity term (w/o $\mathcal{L}_{\text{specific}}$) leads to a drop to 0.7893. These results indicate that both objectives are beneficial and complementary: the consistency loss improves performance by enforcing a view-invariant shared representation, whereas the specificity loss further boosts performance by encouraging view-specific representations to capture complementary discriminative cues beyond the shared representation.

Table 3: Ablation studies for molecular prediction (AUC). **Left table:** Components of our multi-branch prediction head. Specific: the V specific branches only. Concat + Specific: Adds the concatenated branch. **Right table:** Training objectives in our disentanglement module. w/o $\mathcal{L}_{\text{consistency}}$ removes the consistency objective, w/o $\mathcal{L}_{\text{specific}}$ removes the specificity objective.

Datasets	Specific	Concat + Specific	Ours
BCNB-ER	0.9283±0.0177	0.9286±0.0130	**0.9322±0.0136**
BCNB-HER2	0.7703±0.0351	0.7777±0.0183	**0.7894±0.0192**
BCNB-PR	0.8638±0.0326	0.8653±0.0302	**0.8770±0.0287**
BRCA-Molecular	0.7746±0.0193	0.7726±0.0238	**0.7905±0.0285**
BRCA-TP53	0.8516±0.0367	0.8567±0.0307	**0.8654±0.0229**
CRC-KRAS	0.6132±0.0705	0.6307±0.0667	**0.6450±0.0761**
CRC-Molecular	0.8656±0.0281	0.8639±0.0090	**0.8685±0.0272**
SKCM-BRAF	0.5995±0.0712	0.6245±0.0571	**0.6385±0.0641**
Average (Mol)	0.7834	0.7900	**0.8008**

Datasets	w/o $\mathcal{L}_{\text{consistency}}$	w/o $\mathcal{L}_{\text{specific}}$	Ours
BCNB-ER	0.9269±0.0127	0.9214±0.0175	**0.9322±0.0136**
BCNB-HER2	0.7698±0.0144	0.7649±0.0313	**0.7894±0.0192**
BCNB-PR	0.8599±0.0199	0.8670±0.0352	**0.8770±0.0287**
BRCA-Molecular	0.7797±0.0323	0.7829±0.0364	**0.7905±0.0285**
BRCA-TP53	0.8540±0.0366	0.8598±0.0284	**0.8654±0.0229**
CRC-KRAS	0.6396±0.0944	0.6268±0.0650	**0.6450±0.0761**
CRC-Molecular	0.8610±0.0357	0.8669±0.0260	**0.8685±0.0272**
SKCM-BRAF	0.6311±0.0638	0.6246±0.0734	**0.6385±0.0641**
Average (Mol)	0.7902	0.7893	**0.8008**

5 Conclusion

In this work, we revisited the problem of leveraging multiple FMs for WSI analysis. Motivated by the observation that different FMs provide partially shared yet complementary representations and that naive fusion or patch-level selection can be limited, we reformulated FM ensembling as a multi-view multiple instance learning problem. Building on this perspective, we proposed *Multi-View Multiple Instance Learning* (MV-MIL), a plug-and-play framework that treats each FM's features as a distinct view of a WSI, equipped with an Information-Theoretic Disentanglement module to separate shared and view-specific components, and a Multi-Branch Prediction Head to integrate them within standard MIL pipelines. Extensive experiments on 14 public benchmarks covering molecular prediction and survival analysis showed that MV-MIL consistently outperforms both single-FM MIL baselines and representative multi-FM fusion baselines, achieving state-of-the-art results. Ablation studies further confirmed that explicitly disentangling shared and view-specific information, jointly modeling multiple FMs, and instantiating MV-MIL on top of different MIL architectures all contribute to the observed gains.

Acknowledgments

This work was supported by Hong Kong Innovation and Technology Commission (Project No. GHP/006/22GD, PRP/086/24FX and ITC-SKLNSD26SC01) and Shenzhen Science and Technology Innovation Committee Fund (Project No. KCXFZ20230731094059008), HKUST-HKUST(GZ) Cross-Campus Collaborative Research Scheme (Project No. C036), and Guangdong Provincial Department of Science and Technology's 1+1+1 Joint Funding Program for Guangdong-Hong Kong Universities.

References

1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. In: International Conference on Learning Representations (ICLR) (2017)
2. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: Proceedings of the 30th International Conference on Machine Learning (ICML). pp. 1247–1255 (2013)
3. Barber, D., Agakov, F.V.: The im algorithm: A variational approach to information maximization (2003), workshop paper, NIPS 2003 (Information Theory and Learning)
4. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine **25**(8), 1301–1309 (2019). https://doi.org/10.1038/s41591-019-0508-1
5. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
6. Chen, X., Qiu, P., Zhu, W., Wang, H., Li, H., Dong, X., Sun, X., Yu, X., Wang, Y., Razi, A., et al.: Cracking instance jigsaw puzzles: An alternative to multiple instance learning for whole slide image analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21353–21363 (2025)
7. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A.: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nature Medicine **24**, 1559–1567 (2018). https://doi.org/10.1038/s41591-018-0177-5
8. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202 (1972)
9. Federici, M., Dutta, A., Forré, P., Kushman, N., Akata, Z.: Learning robust representations via multi-view information bottleneck. In: 8th International Conference on Learning Representations. OpenReview. net (2020)
10. Gács, P., Korner, J., et al.: Common information is far less than mutual information. Problems of Control and Information Theory **2**, 149–162 (1973)
11. Gao, R., Yang, Z., Yuan, X., Wang, Y., Xia, Y., Zhang, Y., Zheng, B., Gong, Y., Yue, Y., Tu, Y., et al.: Meta-encoder: a unified integration framework for multiple pathological foundation models in cancer detection. Nature Communications (2026)
12. Gustafsson, F.K., Rantalainen, M.: Evaluating computational pathology foundation models for prostate cancer grading under distribution shifts. arXiv preprint arXiv:2410.06723 (2024)
13. Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. Jama **247**(18), 2543–2546 (1982)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)

16. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
17. Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., Grabsch, H.I., Yoshikawa, T., Brenner, H., Chang-Claude, J., Hoffmeister, M., Trautwein, C., Luedde, T.: Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature Medicine **25**, 1054–1056 (2019). https://doi.org/10.1038/s41591-019-0462-y
18. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. Computational and Structural Biotechnology Journal **16**, 34–42 (2018). https://doi.org/10.1016/j.csbj.2018.01.001
19. Kriegeskorte, N., Mur, M., Bandettini, P.A.: Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience **2**, 249 (2008)
20. Lenz, T., Neidlinger, P., Ligero, M., Wölflein, G., van Treeck, M., Kather, J.N.: Unsupervised foundation model-agnostic slide-level representation learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30807–30817 (2025)
21. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2021)
22. Li, J., Hu, J., Sun, Q., Yan, R., Ouyang, M., Guan, T., Han, A., He, C., He, Y.: Can we simplify slide-level fine-tuning of pathology foundation models? arXiv preprint arXiv:2502.20823 (2025)
23. Litjens, G., Kooi, T., Ehteshami Bejnordi, B., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017). https://doi.org/10.1016/j.media.2017.07.005
24. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)
25. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
26. Luo, X., Wang, X., Eweje, F., Zhang, X., Yang, S., Quinton, R., Xiang, J., Li, Y., Ji, Y., Li, Z., et al.: Ensemble learning of foundation models for precision oncology. arXiv preprint arXiv:2508.16085 (2025)
27. Ma, J., Guo, Z., Zhou, F., Wang, Y., Xu, Y., Li, J., Yan, F., Cai, Y., Zhu, Z., Jin, C., et al.: A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. Nature Biomedical Engineering **10**(3), 545–564 (2026)
28. Ma, J., Xu, Y., Zhou, F., Wang, Y., Jin, C., Guo, Z., Wu, J., Tang, O.K., Zhou, H., Wang, X., et al.: Pathbench: A comprehensive comparison benchmark for pathology foundation models towards precision oncology. arXiv preprint arXiv:2505.20202 (2025)
29. Mammadov, A., Le Folgoc, L., Adam, J., Buronfosse, A., Hayem, G., Hocquet, G., Gori, P.: Self-supervision enhances instance-based multiple instance learning methods in digital pathology: a benchmark study. Journal of Medical Imaging **12**(6), 061404–061404 (2025)

30. Neidlinger, P., El Nahhas, O.S., Muti, H.S., Lenz, T., Hoffmeister, M., Brenner, H., van Treeck, M., Langer, R., Dislich, B., Behrens, H.M., et al.: Benchmarking foundation models as feature extractors for weakly supervised computational pathology. Nature biomedical engineering pp. 1–11 (2025)
31. Saillard, C., Jenatton, R., Llinares-López, F., Mariet, Z., Cahané, D., Durand, E., Vert, J.P.: H-optimus-0 (2024), https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0, last accessed 2026-06-30
32. Saillard, C., Jenatton, R., Llinares-López, F., Mariet, Z., Cahané, D., Durand, E., Vert, J.P.: H-optimus-1 (2024), https://huggingface.co/bioptimus/H-optimus-1, last accessed 2026-06-30
33. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
34. Shi, L., Ye, Y., Wang, W., Lei, T., Zhao, Y., Kou, G., Chen, B.: Towards comprehensive information-theoretic multi-view learning. IEEE Transactions on Image Processing (2026)
35. Strouse, D.J., Schwab, D.J.: The deterministic information bottleneck. Neural computation **29**(6), 1611–1630 (2017)
36. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11343–11352 (2024)
37. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv preprint physics/0004057 (2000)
38. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature medicine **30**(10), 2924–2935 (2024)
39. Wang, W., Yan, X., Lee, H., Livescu, K.: Deep variational canonical correlation analysis. arXiv preprint arXiv:1610.03454 (2016), iCLR 2017 submission
40. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)
41. Wen, W., Gong, T., Dong, Y., Yu, S., Dong, B.: Towards the generalization of multi-view learning: An information-theoretical analysis. Information Fusion p. 103776 (2025)
42. Xiao, Y., Hu, Y., Li, B., Zhang, T., Li, Z., Fu, H., Rittscher, J., Yang, K.: Adafusion: Prompt-guided inference with adaptive fusion of pathology foundation models. arXiv preprint arXiv:2508.05084 (2025)
43. Xu, F., Zhu, C., Tang, W., Wang, Y., Zhang, Y., Li, J., Jiang, H., Shi, Z., Liu, J., Jin, M.: Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. Frontiers in oncology **11**, 759007 (2021)
44. Xu, Y., Wang, Y., Zhou, F., Ma, J., Jin, C., Yang, S., Li, J., Zhang, Z., Zhao, C., Zhou, H., et al.: A multimodal knowledge-enhanced whole-slide pathology foundation model. Nature Communications (2025)
45. Xu, Y., Zhang, Z., Zhang, X., Xu, M., Zhou, F., Wang, Y., Ma, J., Xin, Y., Li, D., Lu, C., et al.: A breast vision pathology foundation model for real-world clinical utility. arXiv preprint arXiv:2605.08207 (2026)

46. Yang, S., Wang, Y., Chen, H.: Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In: International conference on medical image computing and computer-assisted intervention. pp. 296–306. Springer (2024)
47. Zeng, Q., Wang, Y., Yang, S., Xu, Y., Zhou, F., Ma, J., Cai, D., Zhang, Z., Qu, L., Wang, Y., et al.: Mambamil+: Modeling long-term contextual patterns for gigapixel whole slide image. arXiv preprint arXiv:2512.17726 (2025)
48. Zhu, W., Chen, X., Qiu, P., Sotiras, A., Razi, A., Wang, Y.: Dgr-mil: Exploring diverse global representation in multiple instance learning for whole slide image classification. In: European conference on computer vision. pp. 333–351. Springer (2024)
49. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)