UNVEILING COLON CANCER THERAPY RESISTANCE: A MULTIPLE-INSTANCE LEARNING APPROACH TO CELL DORMANCY CLASSIFICATION IN HISTOPATHOLOGY IMAGES

December 12, 2024

ABSTRACT

Resistance to cancer therapy and cancer relapse are often driven by a subpopulation of cells that are temporarily arrested in a 'G0-arrest' state [Wiecek et al., 2023]. By employing a weakly-supervised learning pipeline, HistoMIL, developed by [Pan and Secrier, 2023], we benchmarked several multiple-instance learning algorithms to build a robust classifier aiming to predict dormancy in digital pathology slides from colorectal cancer tissue. Through an ensemble of TransMIL models evaluated through 5-fold cross-validation, we obtained a test binary classification performance of AUROC of 0.829 and F1 score of 0.724. We further explored training models to make binary classification through multimodal fusion of clinical features, and regressing G0-arrest scores instead of predicting discrete labels. Throughout our work, we discussed advantages and shortcomings of different MIL algorithms and approaches to prediction, such as trade-offs in classification performance for improved interpretability and alignment with biological expectations. Subsequent interpretability analysis involves heatmap visualization over test colorectal tissue, and this showed clusters of both proliferating and G0-arrest cell populations. We hope this has the potential to assist clinical pathologists in gauging dormancy solely from colorectal H&E stained tissue, serving as a cost-effective alternative to sequencing technologies. The code for our experiments written in HistoMIL is found at `https://github.com/awxlong/HistoMIL`, which we contribute to the computational histopathology research community.

1 Introduction

Colorectal cancer (CRC) ranks as the third most commonly diagnosed cancer, and the second leading cause of cancer associated mortality worldwide [Alboaneen et al., 2023, Sallinger et al., 2023]. It predominantly affects older individuals, with most cases occurring in people aged 50 and above. One of the main drivers of poor survival in patients is post-surgery recurrence. It has been reported that $20 - 50\%$ of patients with CRC will relapse after curative resection [Xiao et al., 2024], with rates varying depending on several factors such as metastatic pattern, tumor anatomical sublocation, and surveyed population [Qaderi et al., 2021, Safari et al., 2023].

'G0-arrest' cells (also referred to as 'dormant') have been garnering the attention of the research community for their role in relapse. When cells exit their normal replicative cycle into a state of G0-arrest, although they might be metabolically active, they cease to grow and have reduced rates of protein synthesis [Cooper, 2000]. Dormant cells are resistant to anti-cancer compounds, such as chemotherapy, that target actively dividing cells. Furthermore, they exhibit immune resistance or adaptation to new environmental niches during metastatic seeding. Altogether, they facilitate minimal residual disease, becoming a major factor associated with cancer relapse. Consequently, [Wiecek et al., 2023] developed through a pan cancer-tissue analysis a transcriptional signature for identifying G0-arrest cells from bulk and single-cell RNA-sequencing data. Thus, monitoring this state in a tumor through sequencing technologies can help study therapeutic resistance.

However, on one hand, bulk-RNA sequencing of cancer tissue is not spatially resolved, and thus obscures the contributions of individual cell types and their interactions within the tumor-microenvironment (TME). On the other-hand, single-cell and spatial transcriptomics (ST) techniques are expensive and are limited in cell coverage compared to

whole-slide images (WSIs) [Levy-Jurgenson et al., 2020]. Furthermore, sequencing technologies, especially spatially-resolved ones, may face several hurdles for routine usage to their novelty, associated costs, and the demand for relevant experienced personnel. We can thus ask whether there exists computational alternatives that can predict both the state of G0-arrest solely from hematoxylin and eosin (H&E) tissue and provide a spatially resolved explanation to such prediction, proving a more accessible alternative than sequencing the tissue.

2 Literature review

Deep learning for molecular-level predictions. In oncology, deep learning (DL) models have demonstrated exceptional capabilities in feature extraction from complex, high-dimensional data like WSIs, thereby enabling precise and timely diagnosis, treatment planning, biomarker identification, localization, (pan-)cancer subtype classification, and prognosis prediction [Song et al., 2023, Tran et al., 2021, Couture, 2022, Lee, 2023]. In particular for analyzing colorectal WSIs, convolutional neural networks (CNNs) have been trained to detect cancer [Alboaneen et al., 2023], classify tumor-immune cells [Parreno-Centeno et al., 2022], distinguish between microsatellite instability (MSI) and microsatellite stable (MSS) subtypes [Hezi et al., 2024], classify homologous recombination deficiency (HRD) and MSI spots [Schirris et al., 2022], detect multiple genetic mutations [Konishi et al., 2023], among others. [El Nahhas et al., 2024] propose a model which predicts the HRD biomarker scores rather than categorical labels of cells in H&E images. They argued that biomarkers of key cancer processes are continuous measurements, and binarizing them result in information loss that may hamper a classifier's performance. Through their experiments, they found that regression significantly enhanced the accuracy of spatially resolved, HRD prediction, and offered a higher prognostic value than classification-based labels. Because both HRD and G0-arrest stage are biomarkers that can be continuous scores, we explore predicting G0-arrest scores in addition to classification. With regards to proliferation biomarkers, [Martino et al., 2024] proposed using conditional adversarial network to identify Ki-67, a protein associated with the G1, S, G2, and M phases of the cell cycle, from H&E images of oral squamous cell carcinoma. A large scale, systematic pan-cancer study by [Arslan et al., 2024] benchmarked 13443 DL models to predict 4481 multiomic biomarkers across 32 cancer types, and they reported high predictive capability of cell proliferation biomarkers, particularly for breast, stomach, colon, and lung cancers, with areas under the receiving operating characteristics (AUROC) reaching up to 0.854. However, to the best of our knowledge, we have yet to find prior work attempting to predict cell dormancy from colorectal WSI, a gap which we aim to fill.

Multiple instance learning. The gigapixel resolution and thus complexity of WSIs present unique computational challenges for the design of a DL pipeline to analyze them. The typical paradigm of pre-processing WSIs consists of tissue segmentation, followed a patch-wise cropping step which divides the gigapixel tissue into thousands of square patches with smaller dimensions, e.g., 224×224 pixels. They are then passed to a feature encoder to obtain a feature representation $\mathcal{W} \in \mathbb{R}^{N \times D}$ of the WSI, where N is the number of patches and D is the dimension of the vector output by the feature encoder. Patch-wise embeddings are aggregated through pooling methods to obtain a global prediction [Tan et al., 2023].

Only slide-level labels are available due to the intense annotation burden associated with WSIs [Tan et al., 2023, Gadermayr and Tschuchnig, 2024]. A WSI is represented as a 'bag' B^n, which is a collection of patches, or 'instances', $\{x_1^n, x_2^n, ..., x_d^n\}$, where each B^n is given single label y^n as follows:

$$y^n = \begin{cases} s & \text{if } \exists j \text{ such that } x_j^n = 1 \\ 0 & \text{if } \forall j, x_j^n = 0 \end{cases} \tag{1}$$

, where $s \in \{0, 1\}$ in a binary classification task, e.g., predicting the presence/absence of G0-arrest cells, or $s \in \mathbb{R}$ if we are predicting a score for the state of G0-arrest. This is, if in certain regions of the tissue G0-arrest cells are identified, then the entire WSI receives a positive label.

Due to this problem setup, we resort to a ***multiple-instance learning*** (MIL) framework, a form of weakly-supervised learning. The goal is to learn to classify slides, as well as the key patches that 'trigger' the slide's label. We train such classifier f by optimizing the negative log likelihood of its parameters θ:

$$-\mathcal{LL}(\mathcal{D}|\theta) = \sum_{i=1}^{N} \ell(y^i, \hat{y}^i) \tag{2}$$

, where $\hat{y}^i = \max_j f(x_j^n)$ can be the max pooling over the N instance embeddings in a bag to determine the bag's label, $\hat{y}^i = \sum_{j=1}^{N} f(x_{ij})$ a sum pooling of all embeddings within the bag, or $\hat{y}^i = \frac{1}{N} \sum_{j=1}^{N} f(x_j)$ can be mean pooling which computes the average of all instance embeddings in the bag (implicitly treating all of them equally), which is not

necessarily the case for WSIs where tumour tissue is more relevant for the task. For each pooling method, the instance embeddings can have attention scores, α, which act as weights representing their relative contribution to the final prediction, e.g., $\hat{y}^i = \frac{1}{N}\sum_{j=1}^{N} \alpha_j f(x_j)$. These attention scores are inherently interpretable as they can be traced back to the original WSI input space, highlighting regions of interest. ℓ is a loss function depending on the output and label modality, which could be the mean squared error in the continuous case, or binary-cross entropy in the discrete case. The choice of architectural backend of f, and the modality of the output (multimodal vs. regression vs. classification) are highly customizable depending on the task specifications and available computational resources.

Foundation models for histopathology. There is an increased interest in the training of foundation models thanks to the massive size and diversity of the training data that is available for representation learning. They are often trained on > 100000 WSIs spanning patient cohorts, cancer tissue types, and across diagnostic tasks. Thanks to such diversity, features output by foundation models are context-dependent, and semantically rich. These can help alleviate the demand for high volumes of data for representation learning, as well as have high prospects of generalization given their pretrained regime across tissue types [Chen et al., 2024]. [Chen et al., 2024] propose UNI, a foundation model based on the Vision Transformer (ViT) pretrained through self-supervised learning using more than 100 million images from over 100000 diagnostic H&E -stained WSIs across 20 major tissue types, including colorectal cancer. [Xu et al., 2024] propose Prov-GigaPath, which employs a scalable variant of the ViT (called LongNet) and is pretrained on 1.3 billion 256×256 pathology image patches in 171189 WSIs spanning 31 major tissue types. Both UNI and Prov-GigaPath are open source, facilitating the integration into our pipeline. Finetuning foundation models is prohibitively expensive, and as such we restrict ourselves in using them as frozen feature extractors in our work.

Multimodal fusion. Consider the following clinical dilemma of a pathologist: after identifying a few morphological abnormalities in a patient's colorectal WSI, they conclude the patient does not need to undergo aggressive chemotherapy. However, would their decision change if they knew the patient was old (> 65) and displayed a high carcinoembryonic antigen (CEA) level? In other words, would their decision change if they based it solely on morphological features versus conditioned jointly on morphological and clinical features? Multimodal fusion is defined as computing a prediction conditioned on a combination of features extracted from different input modalities, such as histological images, genomic data, electronic health records, and a patient's clinical features. The rationale behind is to train a model able to capture cross-modality interactions with the hope of improving the model's predictive expressivity and accuracy [Feng et al., 2024]. [Chen et al., 2022] propose a pan-cancer model integrating WSI with genomic data through late fusion of embeddings to estimate patient survival, elucidating advantages such as mostly outperforming unimodal approaches and improved model explainability thanks to the joint analysis of image and genomic features. Their method is used by [Volinsky-Fremond et al., 2024] to combine tumor stage with endometrial H&E WSI embeddings for predicting recurrence risk. However, we note that multimodal fusion should be carried out carefully to avoid problems such as the incorporation of noisy data that may hamper model performance. Furthermore, in our problem setting, because our G0-arrest labels are computed from RNA-sequencing data, it would be inappropriate to fuse RNA-sequencing data with colorectal WSI to avoid it learning to ignore the morphological features, and instead predict G0-arrest from the RNA-seq features alone, a phenomenon known as 'spurious shortcut' [Lipkova et al., 2022].

HistoMIL. The implementation of a MIL-based pipeline can be cumbersome especially given the unique challenges with handling WSIs and the plethora of MIL algorithms proposed over the years. To facilitate the process of training and evaluating different MIL algorithms tailored to processing cancer WSIs, [Pan and Secrier, 2023] proposed `HistoMIL`, a Python package which encompasses the preprocessing, training, and inference stages of MIL-based pipeline. It leverages the PyTorch Lightning framework to enable efficient and scalable training of MIL models, which consists of techniques like mixed precision training (reducing 32 bits to 16 bits precision), gradient accumulation over batches to reduce the frequency of backpropagation and be able to simulate the processing of larger batches in limited GPU memory, model weight check-pointing which helps resuming failed experiments avoiding re-initializing one from scratch, and logging evaluation metrics to Weights and Biases. `HistoMIL` is also highly customizable with regards to adoption of MIL algorithms. As of writing, the package by default implements ABMIL, DSMIL, and TransMIL algorithms, and assumes the implemented MIL model solves a binary/multiclass classification task. We adapt 8 new MIL algorithms for our benchmark, and implement functions encompassing cross-validation, multimodal fusion, regression, and interpretability analysis.

3 Methodology

Our pipeline is described in Figure 1, which can be broadly split into 3 steps: 1) feature extraction, 2) benchmarking models under 5-fold cross-validation, including ablations and ensembling, to evaluate over the test set, and 3) performing interpretability analysis.

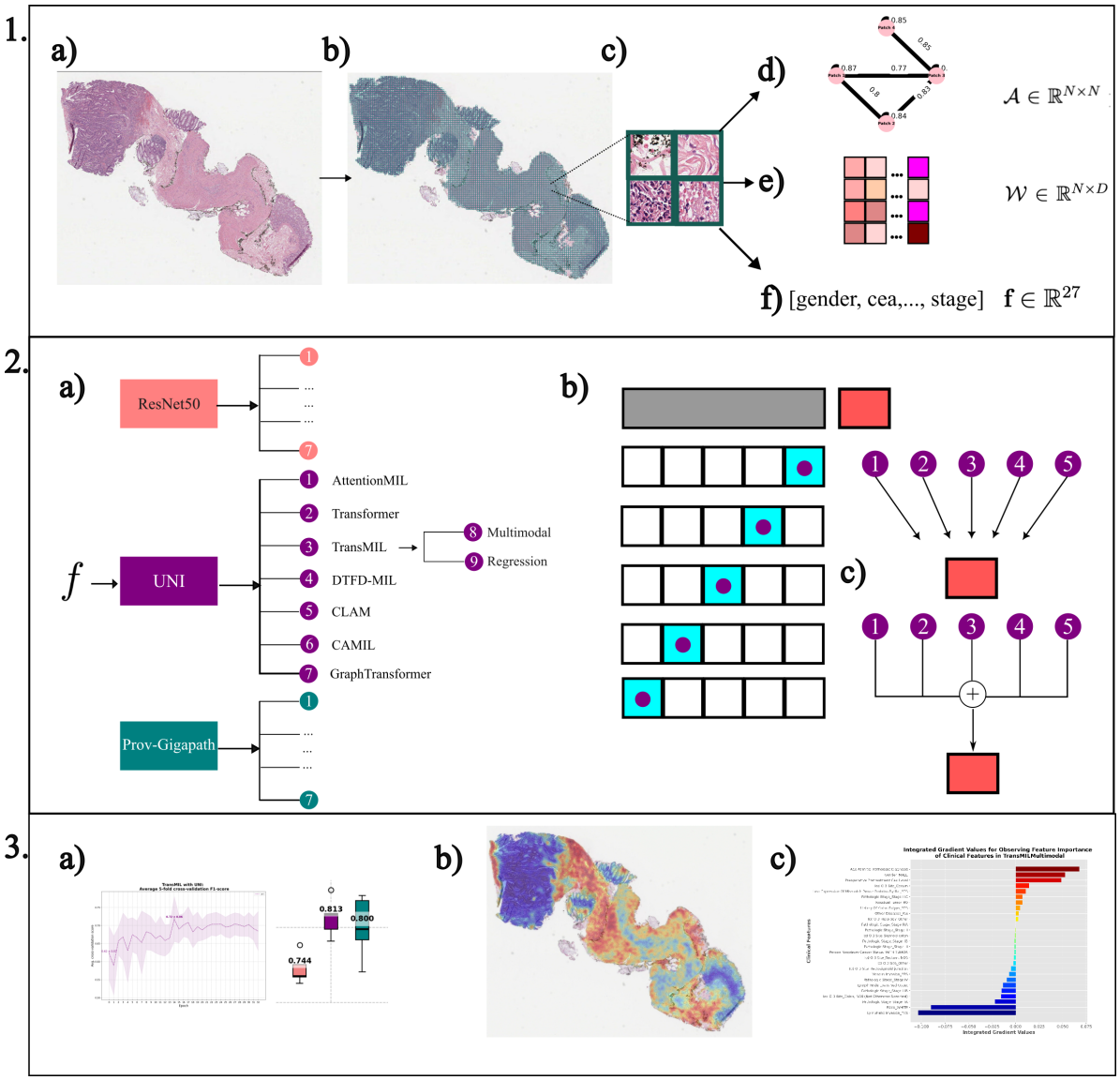

Figure 1: Depiction of our pipeline. First, 1.a) each colorectal WSI 1.b) undergoes tissue segmentation and 1.c) patching. At 1.d), we compute an adjacency matrix $\mathcal{A}$ via Equation 3. For each WSI, 1.e) we compute $\mathcal{W} \in \mathbb{R}^{N \times D}$ using 3 feature encoders: ResNet50, UNI, and Prov-Gigapath. For each patient, 1.f) we extract $\mathbf{f} \in \mathbb{R}^{27}$ clinical features described in Section 3.1. After feature extraction, a classifier f is assembled at 2.a) by implementing each of the MIL algorithms described at Section 3.2 coupled with each of the feature encoders. At 2.b) we depict train-test splitting, along with a 5-fold cross validation framework, which is explained in more detail in Figure 2. 2.c) shows 2 evaluation methods: on one hand we obtain predictions with each fold's optimal model, and on the other hand with an ensemble of the optimal models. Lastly, at 3.a) we report results of our cross-validation and test set. We perform interpretability analysis based on 3.b) heatmap generation, and at 3.c) based on the integrated gradients method for the multimodal model.

3.1 Feature extraction per WSI and patient

We obtain 578 colorectal adenocarcinoma, H&E stained WSIs from the TCGA, each matched with bulk-RNA sequencing data. By employing the genomic signature of [Wiecek et al., 2023], each colon WSI is given a label s (see Equation 1). If it's continuous, s is a score indicating level of quiescence. This score is binarized based on a clinical threshold, whereby if it's negative (≤ 0), $s = 1$ indicating the presence of cells in G0-arrest in the WSI, and if positive (> 0), it represents absence of such.

`HistoMIL` preprocesses one WSI following the segmentation and patching protocol at [Lu et al., 2021], where we choose a patch size of (224×224) with no overlap. We proceed to store a matrix representation $\mathcal{W} \in \mathbb{R}^{N \times D}$ by stacking $D-$dimensional feature vectors computed per each of the N patches of a WSI for each of the following feature encoders: ResNet50 ($D = 2048$), UNI ($D = 1024$) and Prov-Gigapath ($D = 1536$). We do this for each of our WSI, where we note that 1) N is different per slide due to morphologically different tissue per person or anatomical site, and 2) it can range between $[10000, 90000]$. In the interest of training some MIL algorithms with topological constraints, a cropped WSI is represented with an undirected graph $G = (V, E)$ where vertices V correspond to image patches, and $(v_i, v_j) \in E$ are pairwise edges of patches that are adjacent to one another, where in WSIs each patch has at most 8 neighboring patches. G is represented via a weighted adjacency matrix $\mathcal{A} \in \mathbb{R}^{N \times N}$ per WSI, where $\mathcal{A}_{ij} = a_{ij}$ according to the following equation 3:

$$a_{ij} = \begin{cases} \exp\left(-(\mathbf{h}_i - \mathbf{h}_j)^2\right) \text{ iff } (v_i, v_j) \in E, (\mathbf{h}_i, \mathbf{h}_j) \in \mathcal{W} \\ 0, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{3}$$

, where a distance similarity score is computed only if two patches are adjacent to one another, and 0 otherwise. This similarity score is the exponentiated, normalized, Euclidean distance between the feature representations of 2 patches, which injects a bio-topological prior constraint that drives MIL models to attend to patches close and similar to each other [Fourkioti et al., 2024].

Our WSIs belong to 570 unique patients, as for some of them, tissue from multiple anatomical locations was collected. Since we explore multimodal fusion later in our work, we collect the following clinical features $\mathbf{f} \in \mathbb{R}^{27}$ per patient: patient's age, lymph node count, preoperative CEA level, gender, race, other diagnoses, pathological stage, histological site, neoplasm cancer status, venous invasion, lymphatic invasion, history of colon polyps, residual tumor and loss of expression of mismatch repair (MMR). We discuss in detail the selection and preprocessing of the above features at the Appendix C, which involve technicalities such as normalization and mode imputation whilst avoiding train-test leakage, grouping of variables to address class imbalance, one-hot encoding, shadow-based feature selection, among others.

3.2 Benchmarking models under 5-fold CV

We perform a $90\% - 10\%$ train-test split, along with 5-fold cross validation (CV). In contrast to prior work, we employ CV not for hyperparameter tuning nor neural architectural search since that would be prohibitively expensive and cumbersome given our limited GPU cluster resources. Rather, CV is 1) used to get uncertainty estimates of a model's generalization performance, and 2) obtaining independent fold models to build an ensemble for predicting over the test set [El Nahhas et al., 2024] (Figure 2).

At each fold, we benchmark the following MIL algorithms, for each feature encoder. They are chosen based on their novelty, reported efficiency and ease of adoption into the current pipeline in `HistoMIL`:

- **Attention deep MIL (AttentionMIL)**: Proposed by [Ilse et al., 2018], it's a general purpose MIL algorithm that has been used as a baseline in many settings not just restricted to histopathology. It employs the attention mechanism, and assumes permutation invariance of the patches of the slides.
- **Transformer**: transformers have the impressive capabilities of learning through self-attention long-range dependencies and contextualizing concepts in long sequences. In MIL this entails modeling of relationships among instances within a bag, effectively capturing both morphological and spatial information. [Wagner et al., 2023] experimentally show that a fully-transformer based approach results in higher AUROC and generalization performance than pure-CNN or hybrid CNN-Transformer methods to predict biomarkers (MSI, and mutations BRAF, and KRAS) on biopsies of colorectal cancer.
- **Transformer-based MIL (TransMIL)**: Proposed by [Shao et al., 2021], TransMIL alleviates the permutation invariance assumption of the patches in the slide by modelling the correlation amongst instances through a multi-headed attention. The main contrast with the above method is that it replaces the self-attention mechanism with the Nyström attention to reduce the quadratic complexity $O(N^2)$ of the former with a linear complexity of the latter $O(N)$, which is important in our case to deal with slides with up to 90000 patches.

  - With TransMIL, we also explore **multimodal fusion (TransMILMultimodal) of the clinical features** above through late fusion. This consists of passing the embedding of $\mathcal{W}$ and the embedding of the clinical features through an gating-based attention for automatic regularization, followed by the Kronecker product to model for the pairwise feature interactions of the image with clinical modalities before making a final decision [Chen et al., 2020, Volinsky-Fremond et al., 2024].
  - We also explore **regression (TransMILRegression)**, which consists of changing the output of the original TransMIL from a class probability with a range of $[0, 1]$ to a logit with a theoretical range of $[-\infty, +\infty]$. This is accompanied by changing a classification loss function with a regression-based alternative, along with providing G0-arrest ground truth scores instead of binarized labels (see Equation 1)[1].
- **Double-Tier Feature Distillation Multiple Instance Learning (DTFD-MIL)**: Because of our small sample size (< 600), we adopt algorithms designed to address data scarcity. DTFD-MIL [Zhang et al., 2022] address this by partitioning a slide into "pseudo-bags" of patches to virtually increase the number of training bags, and making a slide-level classification decision by aggregating the predictions of the "pseudo-bags", in a process denoted a "double-tier MIL framework".
- **Clustering-constrained Attention Multiple Instance Learning (CLAM)**: Proposed by [Lu et al., 2021], CLAM is also designed to address low-data settings. Through an attention-based mechanism, it learns to focus on the most relevant patch features within a slide by learning to cluster positively from negatively labeled patch features.
- **Context-Aware MIL (CAMIL)**: Proposed by [Fourkioti et al., 2024], CAMIL represents a WSI as a graph and performs "neighbor-constrained attention" to make a classification decision. It consists of injecting the bio-topological constraint stated in Equation 3 to consider the pairwise attention score of patches only if they are adjacent to one another.
- **Graph Transformer**: Proposed by [Zheng et al., 2022], it's a hybrid architecture which also makes use of the graph representation of WSI as CAMIL, whereby the input patches' features go through a vision transformer to make a classification decision.

All the above models, except in TransMILRegression which uses the MSE loss function, are trained by minimizing the binary cross-entropy loss with logits (BCEWithLogits). For all algorithms we train with mixed-precision, a batch size of 1[2], and gradient accumulation over 4 batches to simulate a batch-size of 4, giving us the smoothness and convergence speed of mini-batch optimization. Furthermore, all models, except the Transformer, can complete their 5-fold CV in ≤ 16 GB of GPU memory in less than 3 days. The Transformer is the only which uses an A40 (48 GB of GPU memory), and completes the 5-fold CV regime in less than 6 hours. We reuse the hyperparameters mentioned in each MIL algorithm's paper. For specific details, please see Appendix D.

For evaluation, we report AUROC and the F1-score [Schirris et al., 2022]. For TransMILRegression, where outputs stop being probabilities, we compute instead the Pearson's correlation coefficient (PCC) with ground-truth G0-arrest scores [El Nahhas et al., 2024]. Because we can binarize scores at a clinical threshold of 0, we can compute F1 and compare it across all models benchmarked. Additionally, we also measure performance metrics like validation/test loss, accuracy, precision, specificity, and recall. We also monitor training accuracy and loss to check for training stability and convergence.

3.3 Inference and interpretability analysis

For each classifier, we report the average validation AUROC and F1 across folds per epoch. To prune the exponential increasing space of experiments for us to run, we only choose the best performing algorithm based on the average cross-validation performance, along with its highest performing feature encoder to explore ablation studies: multimodal fusion and regression. This explains why only TransMILMultimodal and TransMILRegression are amongst the benchmarked models above (Figure 2a).

For each classifier, our HistoMIL framework allows us to store the checkpoints at which it achieves the highest validation performance per fold. We evaluate this highest performing model per fold on the test set, and obtain 5 test scores per MIL model for each feature encoder (Figure 2b). Because each highest performing model per fold is an independent model, we further explore whether ensembling them [Khened et al., 2021] by averaging their predictions help improve their generalizability by evaluating them on the test set (Figure 2c).

[1]We only pick TransMIL with the UNI feature encoder to explore multimodal fusion and regression for 2 reasons: it achieves the second highest mean CV F1-score, preceded by the Transformer, and it's affordable to train within 16 GB of GPU memory, unlike the Transformer which requires at least 48 GB.

[2]This is because we can't stack $\mathcal{W}$ as N is different per slide

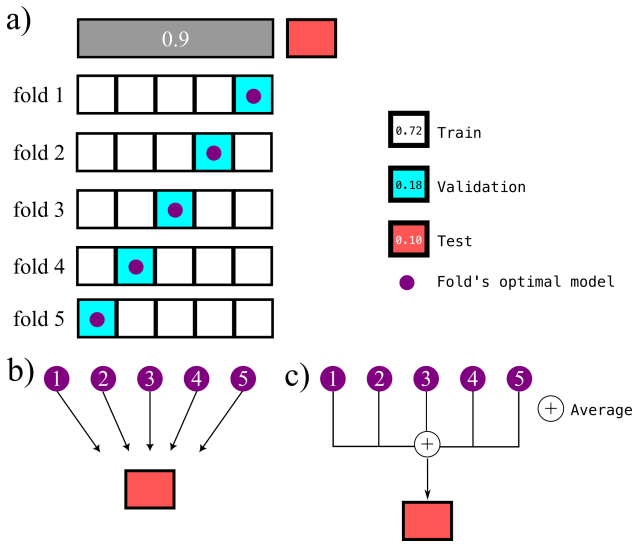

Figure 2: Illustration of our training, CV and evaluation framework with ensembles. a) illustrates train-test splitting, followed by 5-fold CV where our training set is split into 5 equally-sized folds, where 4 are used for training and 1 for validation. Per fold, HistoMIL checkpoints the optimal model by monitoring when it achieves the highest AUROC, where for TransMILRegression it monitors the F1-score. b) Each independent optimal model per fold is evaluated on the test set, such that we are able to obtain uncertainty estimates of the model's generalization capability. c) Afterwards, an ensemble is constructed by averaging the predictions of the 5 optimal models and evaluated on the test set to observe any possible improvement. This pipeline is applied for each MIL algorithm, for each feature encoder, except TransMILMultimodal and TransMILRegression where only UNI is used to avoid an exponential amount of experiments to be run.

Interpretability analysis is done in 2 ways. For all MIL algorithms models except TransMILMultimodal which consists of clinical features, we trace the attention scores back to the original patches they correspond to explain the model output, adopting the method by [Lu et al., 2021][3]. Because in all MIL algorithms, the attention score per patch is pooled to make a prediction (see explanation of Equation 2), visualizing their values over the original patches of the input space helps explain a model's final classification decision or regression score. Since the cell populations in the tissue slide are either in a state of proliferation, or in G0-arrest (i.e. these 2 states are mutually exclusive), patches with high attention scores are regions which drive the model to predict a high likelihood of G0-arrest cells on those patches, while low attention scores correspond to regions unlikely to contain them, i.e., instead there are normal-cycling cells.

For TransMILMultimodal, the model also incorporates clinical features which can not be spatially resolved back to the image input space, and as such the patch-dependent attention scores don't encompass the influence these have over the output. Thus, we further resort to the integrated gradients (IG) method [Sundararajan et al., 2017] to explain which clinical features contribute to the final prediction as done by [Volinsky-Fremond et al., 2024]. IG consists of obtaining the contribution of each input clinical feature to the final prediction by integrating the gradients of the model's output with respect to the input features along a path from a baseline input to the actual input. Such baseline input is a 27^{th}-dimensional zero vector which represents a non-informative state. As a result, IG provides a measure of how much each clinical feature contributes to the prediction compared to a state of null information. A higher, absolute IG value indicates a greater influence of that feature on the final prediction.

All relevant code is found at `https://github.com/awxlong/HistoMIL`, and scripts for running experiments are at `https://github.com/awxlong/scripts_g0_arrest`

[3]While there exists other methods such as GradCAM, we note that they don't fit under a histopathology pipeline since they map outputs back to the original input space. By contrast, in our setting, we don't work with the original WSI due to its gigapixel size, but rather with a feature representation $\mathcal{W}$ of it

4 Results and Discussion

4.1 Foundation feature encoders help with generalization

We benchmark the MIL models and evaluate them on their predictive accuracy on G0-arrest. Our cross-validation results are at Appendix A. Our evaluation over the test set in terms of AUROC (Figure 3) and F1 (Figure 11) suggests that MIL algorithms trained with foundation feature encoders may lead to better generalization performance. We also compute the test F1 over ablations of TransMIL using the UNI feature encoder and report results in Figure 12, where we note that TransMILRegression improves the F1 over the standard TransMIL, while TransMILMultimodal has some performance sacrifice, albeit the latter is compensated with the extensive analysis of its clinical features in Section 4.3. TransMILMultimodal also yields more stable performance across folds, as shown by the tighter uncertainty regions.

We have no prior SoTA results on G0-arrest prediction from histopathological images to compare our current metrics. However, our most performant models can consistently achieve an AUROC greater than 0.75 and F1 greater than 0.65, which underscores the capability of our deep learning models to capture relevant morphological features in the colon tissue to make a binary decision on the presence of G0-arrest cells. This is further explored by the visualization of heatmaps over the tissue by applying our interpretability methods observed in Figure 4. We leave as future work the validation of such heatmaps through ST, and restrict ourselves in highlighting differences of the heatmaps generated across algorithms such as in Figure 7.

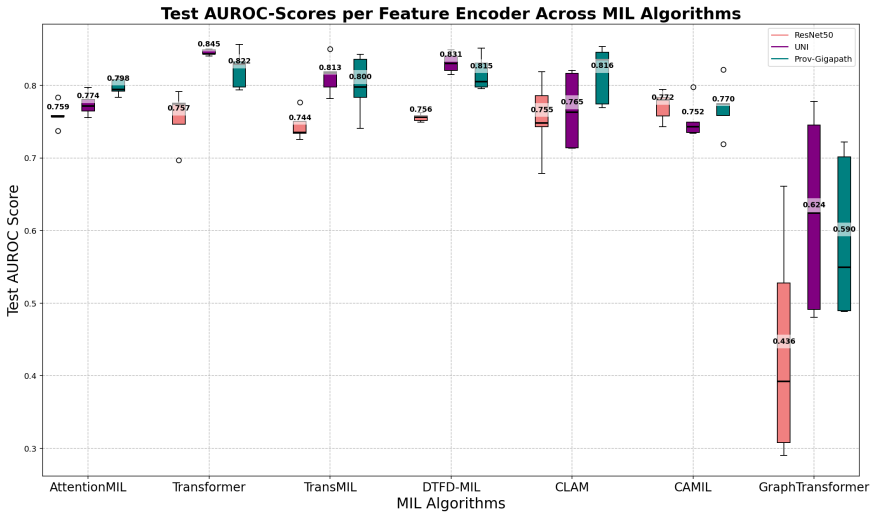

Figure 3: Average test AUROC obtained from the 5 independent optimal models per fold, per feature encoder. We notice that for all algorithms, except CAMIL, at least one of both feature encoders surpasses the performance of the ResNet50 encoder, albeit with overlapping std. errors (i.e. in the figure the purple and teal bars are often higher than their lightcoral counterpart). This suggests that the choice of foundation feature encoders helps with generalization.

4.2 Ensemble modelling improves prediction accuracy

For each MIL algorithm, and for each feature encoder, we construct an ensemble consisting of the 5 optimal models from the CV framework and evaluate on the test set, with results reported at Table 1. We also report an ensemble for each of TransMIL's ablations at Table 2. We generally observe higher scores than in Figures 3, 11, 12, indicating that ensembling helps with generalization performance. Additionally, both UNI and Prov-Gigapath feature encoders' scores are higher than ResNet50, which reinforces the idea that they help with improved model generalizability even when ensembling.

The performance gain from ensembles prompts us to explore how heatmaps generated by an ensemble contrast with those from the single optimal model in cross-validation. Heatmaps from ensembled algorithms consist of averaging the attention scores of each model and plotting them over the WSI. As an example, for TransMIL, we observe how ensembling helps correct a previously wrong prediction made by a single best TransMIL (Figure 5). For a slide labeled 1, the ensemble's heatmap shows more regions of cells in G0-arrest identified, while also attenuating previously very confident regions of cell proliferation. A disadvantage with our ensembles is that none of them provide confidence intervals into their predictions.

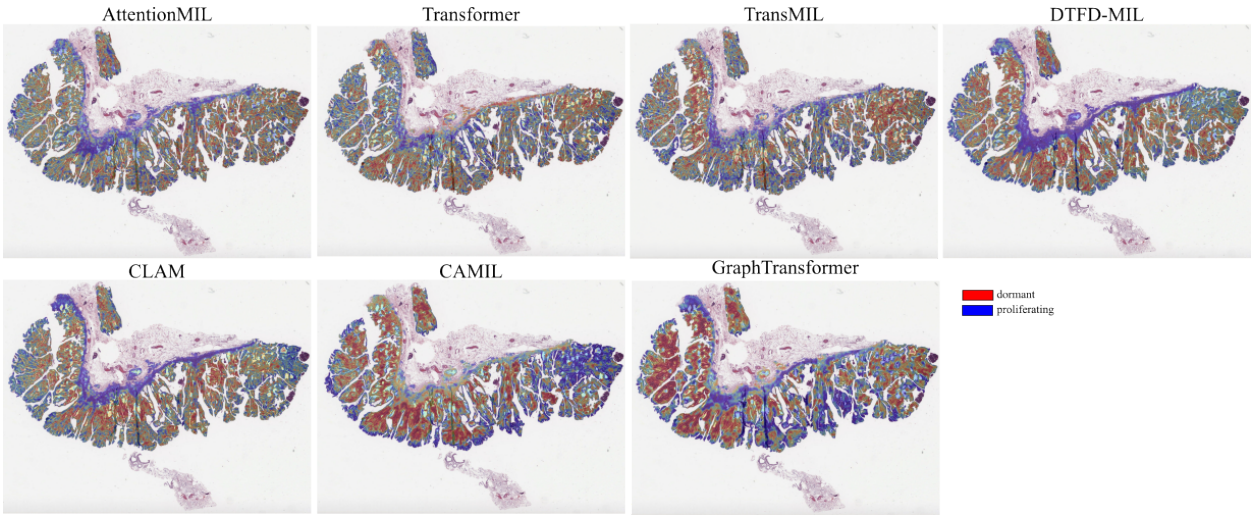

Figure 4: Side-by-side comparison of heatmaps generated by benchmarked algorithms using the UNI feature encoder. Each heatmap is obtained by mapping the average attention scores from an ensemble of the best models per CV fold. All heatmaps explain a TP prediction except for AttentionMIL and GraphTransformer which erroneously make a slide-level prediction of 0. High attention scores correspond regions with high likelihood of cells in G0-arrest, while blue regions has low likelihood of G0-arrest, i.e., cells proliferating by the assumption of mutual exclusivity of classes. Gaussian blur has been applied to avoid a strict demarcation of the patches.

For interested pathologists, we share a more comprehensive view of heatmaps generated by our Ensemble TransMIL in the Appendix Figure 13, spanning those generated in correct and wrong predictions, along with samples of patches where the ensemble bases its predictions on.

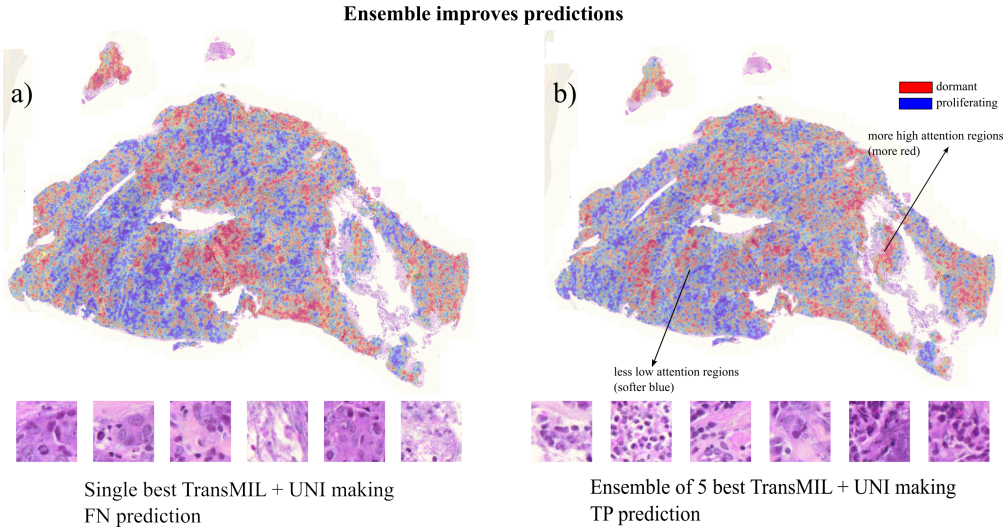

Figure 5: Depiction of how an ensemble improves predictions, and this is reflected in the heatmaps generated. At a) we show a single TransMIL trained with the UNI feature encoder making a false negative prediction on a testpoint. At b), this is corrected into a true positive prediction by an ensemble of the 5 optimal TransMIL according to each of their best validation AUROC achieved per fold. The ensemble is able to identify more regions with high likelihood of G0-arrest cells, while decreasing its belief of the presence of normal-cycling cells in the same regions the single TransMIL believed otherwise. The sampled patches in a) correspond to those with low attention scores, and those in b) are those with high attention scores due to the mutual exclusivity assumption.

	ResNet50		UNI		Prov-Gigapath	
	AUROC	F1	AUROC	F1	AUROC	F1
Transformer	0.759	0.710	***0.859***	***0.780***	0.841	0.737
TransMIL	0.751	0.737	0.829	0.724	0.812	0.750
DTFD-MIL	0.754	0.689	0.831	0.720	0.828	0.741
CAMIL	0.772	0.719	0.779	0.667	0.816	0.746
CLAM	**0.794**	**0.759**	0.776	0.679	**0.844**	**0.750**
AttentionMIL	0.751	0.600	0.779	0.600	0.812	0.654
GraphTransformer	0.325	0.507	0.702	0.667	0.602	0.714

Table 1: Scores obtained from ensemble predictions on the test set. Ensemble consists of the best models per each CV fold which maximized AUROC. In bold we highlight the highest metric across algorithms (column-wise), and in italics we highlight the highest metric across feature encoders (row-wise). This is, the Transformer architecture with the UNI feature encoder achieves the highest test performance.

4.3 Multimodal fusion improves interpretability with some performance sacrifice

We build an Ensemble TransMILMultimodal and analyze the IG values computed over the test set. We observe that a lot of features, particularly categorical ones like ICD-O-3 site and Pathological stages mostly lose their relevance (i.e. average IG value close to 0) for predicting G0-arrest. There is perhaps much heterogeneity regarding these clinical features with respect to predicting the G0-arrest population, which drives the model to base a prediction with morphological features and other clinical features instead. Regardless, we observe that particularly for Pathological Stage IIA (classified under Early Stage), negatively influence a G0-arrest prediction, which could be understood as TransMILMultimodal learning that this stage is associated to the tissue more likely to have populations of proliferating cells. This is consistent with current views of pre-metastasis cancer cell behavior discussing that tumor cell dissemination can occur in the very early stages of disease, long before a tumor is even palpable [Attaran and Bissell, 2021, Lawrence et al., 2023]. On the other hand, TransMILMultimodal identifies Pathological Stage IIIC (Late Stage cancer) as having an average IG value greater than 0, driving the model to predict a high likelihood of G0-arrest populations in the CRC tissue. However, the clinical literature mainly characterizes late stage cancers as consisting of aggressive growth, higher metastatic potential and thus lower survival prospects [Lawrence et al., 2023]; as such they are associated with proliferating cells. Nonetheless, disseminated tumor cells can become dormant in all stages of cancer, and be reactivated due to changes in the tumor microenvironment or therapeutic stress [Truskowski et al., 2023].

By looking at non-zero IG values of relevant clinical features, the Ensemble TransMILMultimodal, identifies patient's age, gender and preoperative CEA level as important features that help understand recurrence through cells in G0-arrest. It's reasonable for both models to focus on preoperative CEA level given its well-established reputation as a prognostic biomarker of CRC, with high CEA levels (> 10 ng/mL) associated with a higher risk of recurrence and metastasis [Lai et al., 2023]. Additionally, age has prompted much research regarding CRC progression and treatment outcomes [Cho et al., 2021]; for instance, age-related biological changes in immune response ('immunosenescence') [Thoma et al., 2021] leads to older patients being associated with higher prevalence of senescent T cells, which are less effective at responding to tumors. Furthermore, research has corroborated the existence of sexual dimorphisms with regards to CRC response to treatment efficacy or toxicity [Baraibar et al., 2023], or survival advantages [Geddes et al., 2022], which could be partly explained by an interplay of senescent and proliferating cells.

On the other hand, negative IG values correspond to features contributing to a prediction of proliferating cells related to recurrence. Research has identified racial disparities regarding recurrence incidence, with [Snyder et al., 2020] finding that amongst US patients with locoregional CRC, black patients experience a higher risk of recurrence and mortality compared to white patients. In addition, features like lymphatic invasion indicates tumor proliferation to regional lymph nodes and distant sites. Lymphatic invasion, nonetheless, has also been associated with cancer recurrence through tumor dormancy. This is because cancer cells which enter lymphatic vessels and colonize lymph nodes can remain dormant there for extended periods, leading to relapse [Giancotti, 2013]. We note that a limitation with both interpretability methods (feature importance and IG) is that neither shows for which particular values or range of values of each feature contribute to the final G0-arrest classification decision, nor in which direction do they push/pull the decision boundary. As such, we can at most state that these features are relevant, but can't stratify G0-arrest based on the values of these features.

The literature on the understanding of CRC recurrence is nuanced and multi-faceted, and generally it's inconclusive whether it's driven mainly through tumor proliferation or reactivation of dormant tumor cells. Our heatmaps and multimodal analysis can aid clinical pathologists in navigating through this complicated tumor landscape.

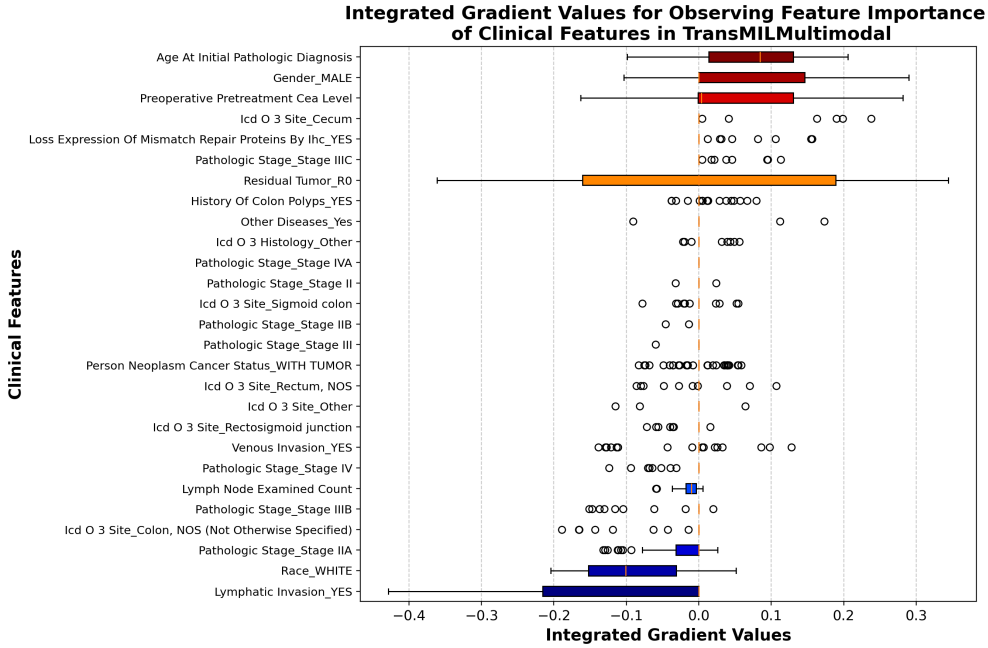

Figure 6: Descendently ranked IG values per clinical feature obtained by averaging the IG values obtained with an ensemble TransMILMultimodal predicting over the test set. We use the same color code as before, except that with IG values, the theoretical range extends to $[-\infty, +\infty]$, where positive IG values refer to features contributing to a positive prediction, while negative ones to a negative prediction. IG values of 0 indicate the corresponding features provide no significant information to make a prediction compared to a null baseline of 0.

	AUROC	F1	PCC
TransMIL	0.829	0.724	N/A
+ Multimodal	0.805	0.679	N/A
+ Regression	N/A	**0.786**	0.312

Table 2: Scores obtained from ensemble predictions on the test set. Ensemble consists of the best models per each CV fold which maximized AUROC. For multimodal and regression, we only perform experiments on TransMIL using the UNI feature encoder. For regression, we note that only PCC is available to measure the correlation of the continuous predictions with the G0-arrest scores. F1 is measured via binarizing the regression scores with a clinical threshold of 0 and comparing with the binary ground truth labels. The first row is the same as in Table 1. We obtain the highest F1 through binarizing regression scores.

4.4 Inductive biases yield more biologically meaningful predictions

Spatial context-awareness From the heatmaps shown in Figure 4, we observe that the spatial constraints of CAMIL and GraphTransformer enable visualizing more pronounced clusters of cell populations, while for alternatives the cell populations are more scattered. Despite this comes at some performance sacrifice, where CAMIL is our fourth performant model and the GraphTransformer is amongst the worst, the heatmaps indicate their local predictions align closer to biological expectations regarding both the proliferating and quiescent cells to cluster with each other (see a closer look at Figure 7). The drop in performance could be explained as follows: if individual patches were misidentified to contain G0-arrest cells, then subsequent patches would also be considered to erroneously contain G0-arrest cells due to the adjacency constraints that make neighboring patches influence each other.

Biological continuum awareness Additionally, predicting G0-arrest scores instead of dichotomized labels is more biologically plausible. Even though there is a clear demarcation regarding cell states between G0-arrest, and proliferating cells, the cell cycle itself is a spectrum, and cells could be in a state transitioning to G0-arrest or exiting. As such, binarizing G0-arrest scores could lead to information loss [El Nahhas et al., 2024] regarding this biological spectrum. While our Ensemble TransMILRegression results show poor PCC with regards to ground truth scores (Table 2), interestingly, if we train the model through regression and binarize the output scores, Ensemble TransMILRe-

gression achieves the highest test F1 (0.786) amongst all the models benchmarked. This underlies the advantages of learning to predict regression scores helping the model become more expressive and improve accuracy of prediction. [El Nahhas et al., 2024] also argue that regression-based models yield heatmaps highlighting more clinically relevant regions. Whilst we compare heatmaps amongst TransMIL ablations in Figure 14, we note that due to our lack of ground truth annotations at a patch-level regarding the populations of cells in G0-arrest, we are unable to comment on the biological fidelity of regression-based heatmaps. We thus leave this as future work.

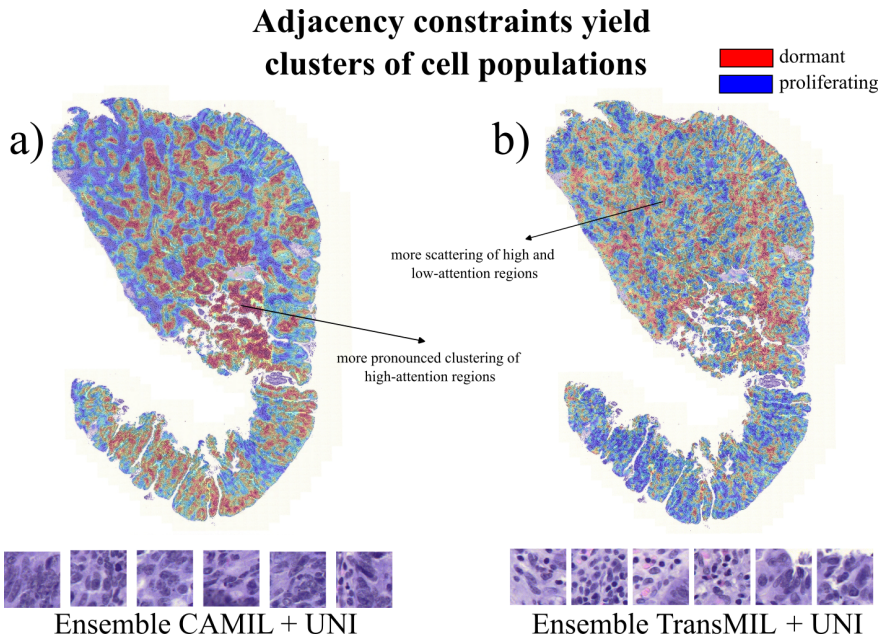

Figure 7: Adjacency constraints introduced via the graph representation of a WSI helps the model visualize more pronounced clusters of cell populations. While this comes at the expense of some performance loss, the spatially-constrained heatmaps produced by a) CAMIL and GraphTransformer align more with biological expectations regarding both the proliferating and quiescent cells to cluster with each other. This is in contrast to b) TransMIL and other algorithms which tend to produce heatmaps with more scattered cell populations. At the bottom of each heatmap we show a sample of 6 patches with the highest attention scores contributing to the TP prediction.

4.5 Limitations and future work

G0-arrest and tumor heterogeneity. Our main interest revolves around guiding therapy to be effective against CRC tissue with populations of G0-arrest tumor cells. However, we note that our slide-level labels y are computed from bulk-RNA sequencing data, thus there is a mix of genomic signals derived from the TME which is not unique to tumor cells, but also from a mixture of fibroblasts, immune and endothelial cells. Future work can exploit ST at a single-cell resolution to demarcate tumor and somatic cells in G0-arrest, however, they could prove relatively inaccessible due to their expensive costs.

Out-of-distribution evaluation. We validate our models through 5-fold cross-validation coupled with a TCGA in-domain test-split. Future work can explore the generalization capability of each algorithm through test-splits stratified by different clinical sites [El Nahhas et al., 2024] or datasets belonging to patient cohorts different to TCGA, such as those sourced from the Yorkshire Cancer Research Bowel Cancer Improvement Programme (YCR-BCIP) [Wagner et al., 2023], to evaluate out-of-distribution generalization. Furthermore, regarding interpretability analysis, we aim to employ CRC tissue which has undergone ST analysis in order for us to evaluate the accuracy of the heatmaps produced. Such evaluation would help us answer questions such as whether the use of foundation feature encoders (and with which algorithms) highlight more biologically-relevant important regions in the colon WSI, in addition to their slide-level prediction accuracy.

Graph theory. Similar to [Parreno-Centeno et al., 2022], we can also resort to graph theory to analyse the cell-cell interactions over a CRC tissue. This method consists of employing nuclei segmentation tools like CellVIT [Horst et al., 2024] or CPP-Net [Chen et al., 2023] over the CRC WSI to build a cell-cell interaction graph. We can

then query this graph through knowledge bases like Neo4J to unravel tumour-immune cell dependencies that could be exploited therapeutically. Thus, this would add an additional layer of interpretability analysis to our pipeline, which would prove beneficial for guiding therapy.

Pan-cancer modelling Another direction of research worth exploring is predicting the G0-arrest state across cancer tissues [Arslan et al., 2024]. We hypothesize that in this cross-tissue setting, the benefits of employing foundation feature encoders like UNI would be more pronounced compared to our current setting where we only work with CRC tissue. This is because the embeddings provided by foundation models are semantically rich given their representational learning over massive amounts of cross-tumoral tissue, which aid in generalization better than standard feature encoders like ResNet50, and cheaper to use if compared to custom training a feature encoder. This would greatly increase the size and heterogeneity of our dataset, which allows us to perform more thorough evaluation, but also introduce new challenges since the G0-arrest signature varies by tissue.

5 Conclusion

Our comprehensive benchmarking allows us to look back to our original research aims and confirm deep learning can gauge the G0-arrest population solely from H&E CRC tissue. Ensembling CV models, using foundation feature encoders, multimodal fusion of clinical features, introduction of spatial inductive biases and regression score prediction bring advantages and disadvantages regarding the model's predictive performance and elucidation of the model's internal mechanisms for making a decision. Ensembling and using foundation feature encoders generally provide improved generalization. The fusion of clinical features slightly hampered test classification performance, but enabled a thorough discussion of clinical features in the context of studying G0-arrest and relapse. Generated heatmaps provide interpretable results regarding the spatial composition of G0-arrest cells, and graph-based constraints drive heatmaps to be more biologically plausible reflected by more pronounced clusters of cell populations.

We also contribute to the computational histopathology community with our MIL pipeline, HistoMIL, to advance cancer research, benchmarking and analysis. There is much work to explore, such as cross-tumoral tissue classification of G0-arrest. We are intrigued to observe how deep learning can be further used to aid pathologists with understanding the evolution of the tumor landscape. For reference, we release all our code (including data analysis, plots, scripts for running experiments, among others) for executing our pipeline at `https://github.com/awxlong/HistoMIL`

References

[Alboaneen et al., 2023] Alboaneen, D., Alqarni, R., Alqahtani, S., Alrashidi, M., Alhuda, R., Alyahyan, E., and Alshammari, T. (2023). Predicting colorectal cancer using machine and deep learning algorithms: Challenges and opportunities. *Big Data and Cognitive Computing*, 7:74.

[Arslan et al., 2024] Arslan, S., Schmidt, J., Bass, C., Mehrotra, D., Geraldes, A., Singhal, S., Hense, J., Li, X., Pandu, R.-L., Maiques, O., Nikolas Kather, J., and Pandya, P. (2024). A systematic pan-cancer study on deep learning-based prediction of multi-omic biomarkers from routine pathology images. *Communications Medicine*, 4.

[Attaran and Bissell, 2021] Attaran, S. and Bissell, M. J. (2021). The role of tumor microenvironment and exosomes in dormancy and relapse. *Seminars in Cancer Biology*.

[Baraibar et al., 2023] Baraibar, I., Ros, J., Saoudi, N., Salva, F., García, A., Castells, M. R., Tabernero, J., and Elez, E. (2023). Sex and gender perspectives in colorectal cancer. *ESMO Open*, 8:101204.

[Chen et al., 2024] Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L. L., Wang, J. J., Vaidya, A., Le, L. P., Gerber, G., Sahai, S., Williams, W., and Mahmood, F. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, pages 1–13.

[Chen et al., 2020] Chen, R. J., Lu, M. Y., Wang, J., Williamson, D. F. K., Rodig, S. J., Lindeman, N. I., and Mahmood, F. (2020). Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, pages 1–1.

[Chen et al., 2022] Chen, R. J., Lu, M. Y., Williamson, D. F., Chen, T. Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., and Mahmood, F. (2022). Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40:865–878.e6.

[Chen et al., 2023] Chen, S., Ding, C., Liu, M., and Tao, D. (2023). Cpp-net: Context-aware polygon proposal network for nucleus segmentation. *IEEE Transactions on Image Processing*, 32:980–994.

[Cho et al., 2021] Cho, M. Y., Siegel, D. A., Demb, J., Richardson, L. C., and Gupta, S. (2021). Increasing colorectal cancer incidence before and after age 50: Implications for screening initiation and promotion of "on-time" screening. *Digestive Diseases and Sciences*, 67:4086–4091.

[Cooper, 2000] Cooper, G. M. (2000). The eukaryotic cell cycle.

[Couture, 2022] Couture, H. D. (2022). Deep learning-based prediction of molecular tumor biomarkers from h&e: A practical review. *Journal of Personalized Medicine*, 12:2022.

[El Nahhas et al., 2024] El Nahhas, O., Chiara, L., Carrero, Z. I., Treeck, M. v., Kolbinger, F. R., Hewitt, K. J., Muti, H. S., Graziani, M., Zeng, Q., Calderaro, J., Ortiz-Brüchle, N., Yuan, T., Hoffmeister, M., Brenner, H., Brobeil, A., Reis-Filho, J. S., and Nikolas Kather, J. (2024). Regression-based deep-learning predicts molecular biomarkers from pathology slides. *Nature Communications*, 15.

[Feng et al., 2024] Feng, X., Shu, W., Li, M., Li, J., Xu, J., and He, M. (2024). Pathogenomics for accurate diagnosis, treatment, prognosis of oncology: a cutting edge overview. *Journal of translational medicine*, 22.

[Fourkioti et al., 2024] Fourkioti, O., De Vries, M., and Bakal, C. (2024). Camil: Context-aware multiple instance learning for cancer detection and subtyping in whole slide images. *The Twelfth International Conference on Learning Representations*.

[Gadermayr and Tschuchnig, 2024] Gadermayr, M. and Tschuchnig, M. (2024). Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Computerized Medical Imaging and Graphics*, pages 102337–102337.

[Geddes et al., 2022] Geddes, A. E., Ray, A. L., Nofchissey, R. A., Esmaeili, A., Saunders, A., Bender, D. E., Khan, M., Sheeja, A., Ahrendsen, J. T., Li, M., Fung, K.-M., Jayaraman, M., Yang, J., Booth, K. K., Dunn, G. D., Carter, S. N., and Morris, K. T. (2022). An analysis of sexual dimorphism in the tumor microenvironment of colorectal cancer. *Frontiers in Oncology*, 12.

[Giancotti, 2013] Giancotti, F. (2013). Mechanisms governing metastatic dormancy and reactivation. *Cell*, 155:750–764.

[Hezi et al., 2024] Hezi, H., Gelber, M., Balabanov, A., Yosef E, M., and Freiman, M. (2024). Cimil-crc: a clinically-informed multiple instance learning framework for patient-level colorectal cancer molecular subtypes classification from h&e stained images. *arXiv (Cornell University)*.

[Horst et al., 2024] Horst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., Ugurel, S., Siveke, J., Grünwald, B., Egger, J., and Kleesiek, J. (2024). Cellvit: Vision transformers for precise cell segmentation and classification. *Medical image analysis*, 94:103143–103143.

[Ilse et al., 2018] Ilse, M., Tomczak, J., and Welling, M. (2018). Attention-based deep multiple instance learning. *proceedings.mlr.press*, pages 2127–2136.

[Khened et al., 2021] Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G., and Srinivasan, B. (2021). A generalized deep learning framework for whole-slide image segmentation and analysis. *Scientific Reports*, 11.

[Konishi et al., 2023] Konishi, T., Grynkiewicz, M., Saito, K., Kobayashi, T., Goto, A., Umakoshi, M., Iwata, T., Nishio, H., Katoh, Y., Fujita, T., Matsui, T., Sugawara, M., and Sano, H. (2023). Deep learning-based approach to predict multiple genetic mutations in colorectal and lung cancer tissues using hematoxylin and eosin-stained whole-slide images. *Journal of Clinical Oncology*, 41:1549–1549.

[Lai et al., 2023] Lai, Y.-H., Chang, Y.-T., Chang, Y.-J., Tsai, J.-T., Li, M.-H., and Lin, J.-C. (2023). Predictive value of the interaction between cea and hemoglobin in neoadjuvant ccrt outcomes in rectal cancer patients. *Journal of Clinical Medicine*, 12:7690–7690.

[Lawrence et al., 2023] Lawrence, R., Watters, M., Davies, C. R., Pantel, K., and Lu, Y.-J. (2023). Circulating tumour cells for early detection of clinically relevant cancer. *Nature Reviews Clinical Oncology*, 20:487–500.

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

[Levy-Jurgenson et al., 2020] Levy-Jurgenson, A., Tekpli, X., Kristensen, V. N., and Yakhini, Z. (2020). Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. *Scientific Reports*, 10.

[Lipkova et al., 2022] Lipkova, J., Chen, R. J., Chen, B., Lu, M. Y., Barbieri, M., Shao, D., Vaidya, A. J., Chen, C., Zhuang, L., Williamson, D. F. K., Shaban, M., Chen, T. Y., and Mahmood, F. (2022). Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*, 40:1095 – 1110.

[Lu et al., 2021] Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555–570.

[Martino et al., 2024] Martino, F., Ilardi, G., Varricchio, S., Russo, D., Maria, R., Staibano, S., and Merolla, F. (2024). A deep learning model to predict ki-67 positivity in oral squamous cell carcinoma. *Journal of Pathology Informatics*, 15:100354–100354.

[Messenger et al., 2012] Messenger, D. E., Driman, D. K., and Kirsch, R. (2012). Developments in the assessment of venous invasion in colorectal cancer: implications for future practice and patient outcome. *Human Pathology*, 43:965–973.

[Nadorvari et al., 2024] Nadorvari, M. L., Lotz, G., Kulka, J., Kiss, A., and Timar, J. (2024). Microsatellite instability and mismatch repair protein deficiency: equal predictive markers? *Pathology & Oncology Research*, 30.

[Pan and Secrier, 2023] Pan, S. and Secrier, M. (2023). Histomil: A python package for training multiple instance learning models on histopathology slides. *iScience*, 26:108073–108073.

[Parreno-Centeno et al., 2022] Parreno-Centeno, M., Malagoli Tagliazucchi, G., Withnell, E., Pan, S., and Secrier, M. (2022). A deep learning and graph-based approach to characterise the immunological landscape and spatial architecture of colon cancer tissue. *bioRxiv (Cold Spring Harbor Laboratory)*.

[Qaderi et al., 2021] Qaderi, S. M., Galjart, B., Verhoef, C., Slooter, G. D., Koopman, M., Verhoeven, R. H. A., de Wilt, J. H. W., and van Erning, F. N. (2021). Disease recurrence after colorectal cancer surgery in the modern era: a population-based study. *International Journal of Colorectal Disease*, 36:2399–2410.

[Safari et al., 2023] Safari, M., Mahmoudi, L., Baker, E. K., Roshanaei, G., Fallah, R., Shahnavaz, A., and Asghari-Jafarabadi, M. (2023). Recurrence and postoperative death in patients with colorectal cancer: A new perspective via semi-competing risk framework. *PubMed Central*, 34:736–746.

[Sallinger et al., 2023] Sallinger, K., Gruber, M., Müller, C.-T., Bonstingl, L., Pritz, E., Pankratz, K., Gerger, A., Smolle, M. A., Aigelsreiter, A., Surova, O., Svedlund, J., Nilsson, M., Kroneis, T., and El-Heliebi, A. (2023). Spatial tumour gene signature discriminates neoplastic from non-neoplastic compartments in colon cancer: unravelling predictive biomarkers for relapse. *Journal of translational medicine*, 21.

[Schirris et al., 2022] Schirris, Y., Gavves, E., Nederlof, I., Horlings, H. M., and Teuwen, J. (2022). Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from h&e whole-slide images in colorectal and breast cancer. *Medical Image Analysis*, 79:102464.

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

[Snyder et al., 2020] Snyder, R. A., Hu, C.-Y., Zafar, S. N., Francescatti, A., and Chang, G. J. (2020). Racial disparities in recurrence and overall survival in patients with locoregional colorectal cancer. *JNCI: Journal of the National Cancer Institute*.

[Song et al., 2023] Song, A. H., Jaume, G., Williamson, D., Lu, M., Vaidya, A., Miller, T. R., and Mahmood, F. (2023). Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*.

[Sundararajan et al., 2017] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 70:3319–3328.

[Tan et al., 2023] Tan, L., Li, H., Yu, J., Zhou, H., Wang, Z., Niu, Z., Li, J., and Li, Z. (2023). Colorectal cancer lymph node metastasis prediction with weakly supervised transformer-based multi-instance learning. *Medical & Biological Engineering & Computing*, 61:1565–1580.

[Thoma et al., 2021] Thoma, O.-M., Neurath, M. F., and Waldner, M. J. (2021). T cell aging in patients with colorectal cancer - what do we know so far? *Cancers*, 13:6227.

[Tran et al., 2021] Tran, K. A., Kondrashova, O., Bradley, A., Williams, E. D., Pearson, J. V., and Waddell, N. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine*, 13.

[Truskowski et al., 2023] Truskowski, K., Amend, S. R., and Pienta, K. J. (2023). Dormant cancer cells: programmed quiescence, senescence, or both? *Cancer and Metastasis Reviews*, 42:37–47.

[Volinsky-Fremond et al., 2024] Volinsky-Fremond, S., Horeweg, N., Andani, S., Barkey Wolf, J., Lafarge, M. W., de Kroon, C. D., Ortoft, G., Hogdall, E., Dijkstra, J., Jobsen, J. J., Lutgens, L. C. H. W., Powell, M. E., Mileshkin, L. R., Mackay, H., Leary, A., Katsaros, D., Nijman, H. W., de Boer, S. M., Nout, R. A., de Bruyn, M., Church, D., Smit, V. T. H. B. M., Creutzberg, C. L., Koelzer, V. H., and Bosse, T. (2024). Prediction of recurrence risk in endometrial cancer with multimodal deep learning. *Nature Medicine*, pages 1–12.

[Wagner et al., 2023] Wagner, S. J., Reisenbüchler, D., West, N. P., Moritz Niehues, J., Zhu, J., Foersch, S., Patrick Veldhuizen, G., Quirke, P., Grabsch, H., , v., Hutchins, G., Richman, S. D., Yuan, T., Langer, R., Jenniskens, J. C. A., Offermans, K., Mueller, W., Gray, R., Gruber, S. B., Greenson, J. K., Rennert, G., Bonner, J. D., Schmolze, D., Jonnagaddala, J., Hawkins, N. J., Ward, R. L., Morton, D., Seymour, M., Magill, L., Nowak, M., Hay, J., Koelzer, V. H., Church, D. N., Matek, C., Geppert, C., Peng, C., Zhi, C., Ouyang, X., James, J., Loughrey, M. B., Salto-Tellez, M., Brenner, H., Hoffmeister, M., Truhn, D., Schnabel, J. A., Boxberg, M., Peng, T., Nikolas Kather, J., Church, D. N., Domingo, E., Edwards, J., Glimelius, B., Gögenür, I., Harkin, A., Hay, J., Iveson, T., Jaeger, E., Kelly, C., Kerr, R., Maka, N., Morgan, H., Oien, K. A., Orange, C., Palles, C., Roxburgh, C. S., Sansom, O. J., Saunders, M., and Tomlinson, I. (2023). Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell*, 41:1650–1661.e4.

[Wiecek et al., 2023] Wiecek, A. J., Cutty, S. J., Kornai, D., Parreno-Centeno, M., Gourmet, L. E., Malagoli Tagliazucchi, G., Jacobson, D. H., Zhang, P., Xiong, L., Bond, G. L., Barr, A. R., and Secrier, M. (2023). Genomic hallmarks and therapeutic implications of g0 cell cycle arrest in cancer. *Genome Biology*, 24.

[Xiao et al., 2024] Xiao, H., Weng, Z., Sun, K., Shen, J., Lin, J., Chen, S., Li, B., Shi, Y., Kuang, M., Song, X., Weng, W., and Peng, S. (2024). Predicting 5-year recurrence risk in colorectal cancer: development and validation of a histology-based deep learning approach. *British Journal of Cancer*, 130:951–960.

[Xu et al., 2024] Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, T., Lee, S., Weerasinghe, R., Wright, B. J., Robicsek, A., Piening, B., Bifulco, C., Wang, S., and Poon, H. (2024). A whole-slide foundation model for digital pathology from real-world data. *Nature*, pages 1–8.

[Zhang et al., 2022] Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S. E., and Zheng, Y. (2022). Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Zheng et al., 2022] Zheng, Y., Gindra, R., Green, E. J., Burks, E., Betke, M., Beane, J., and Kolachalama, V. B. (2022). A graph-transformer for whole slide image classification. *IEEE Transactions on Medical Imaging*, 41:3003–3015.

[Zukic, 2024] Zukic, M. (2024). Predict schizophrenia using brain anatomy. *Applied AI (COMP0189) coursework*.

A 5-fold cross-validation results

Our cross-validation (CV) results for each feature encoder and MIL algorithm are shown at Figure 8 for the AUROC metric and at Figure 9 for the F1. Uncertainty regions correspond to the standard deviations of the metric averaged across folds, and these are spread across epochs. We notice much overlap amongst the regions of different feature encoders, which indicates that during cross-validation, the use of foundation feature encoders didn't show much performance improvement.

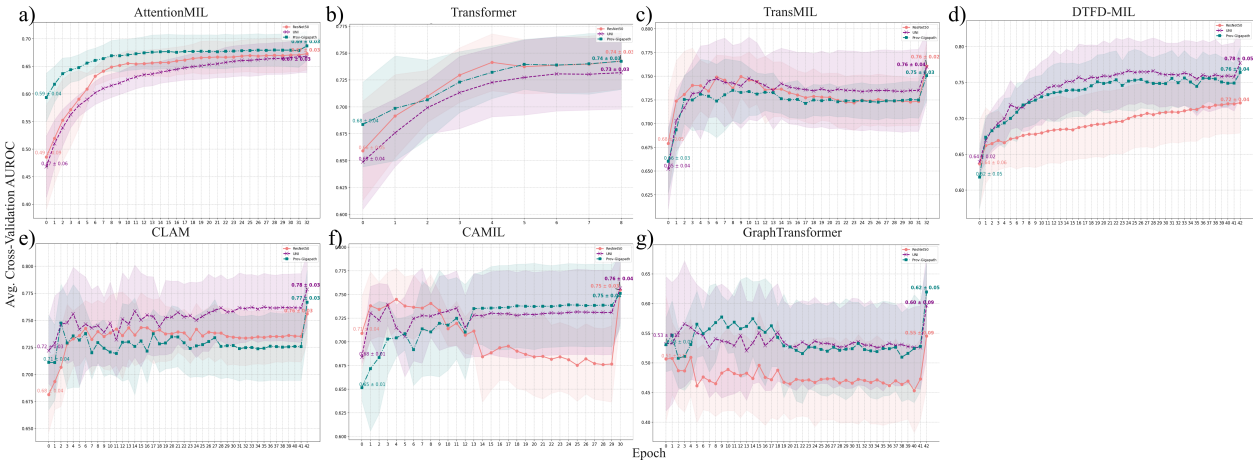

Figure 8: Average AUROC across folds per epoch shown per classifier. We label two "milestones", which is the average performance at the beginning of training, and in bold we show the highest mean cross-validation AUROC achieved at the end of training to illustrate the improvement brought by learning. There is much overlap in CV AUROC's uncertainty regions, with occasional noticeable demarcation such as in d) where the ResNet50 encoder consistently yields lower performance across epochs than its foundation model alternatives.

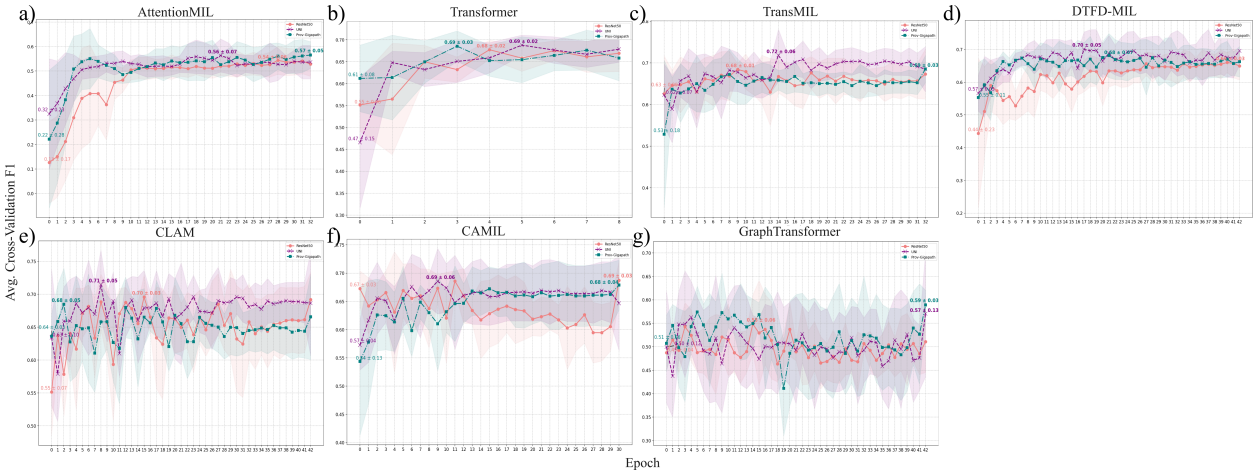

Figure 9: Average F1 across folds per epoch shown per classifier. We label two "milestones" in the same manner as in Figure 8, where we observe that the highest mean cross-validation F1 is not necessarily achieved at the end of training.

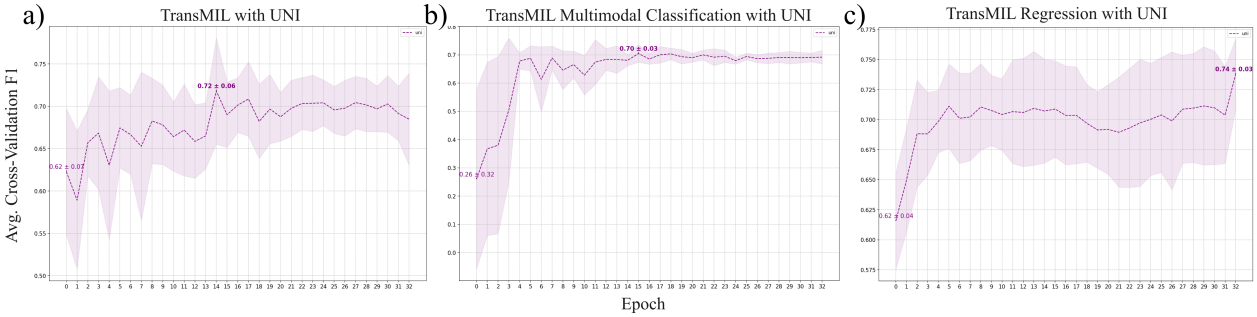

Figure 10: Average F1 across folds per epoch shown for ablations of TransMIL with UNI: TransMILMultimodal and TransMILRegression. We label two "milestones" in the same manner as in Figure 8. a) is the same lineplot as Figure 9c's UNI encoder. b) Interestingly, the average scores across folds is more stable, suggesting that multimodal fusion stabilizes training across folds.

Our CV results help guide how we further explore multimodal fusion and regression by pruning the space of all possible experiments to run, i.e., we avoid exhaustive ablation exploring multimodal fusion with all MIL algorithms and feature encoders. From the plots, we generally observe that classifier consisting of the Prov-Gigapath and UNI feature encoders have slightly higher mean performance than ResNet50. In addition, TransMIL is the one which achieves amongst the highest CV AUROC (≈ 0.75) and highest mean CV F1-score (0.72 ± 0.06) (albeit it's closely followed by CLAM and DTFD-MIL). Because of this, we explore multimodal fusion of clinical features and outputting regression scores only with TransMIL with the UNI feature encoder.

We only show the mean CV F1 across folds in Figure 10 because PCC is not available for the base TransMIL and TransMILMultimodal, while AUROC is not available for TransMILRegression. In this regard, F1 provides a common score to compare ablations of TransMIL.

B Test results

C Clinical feature selection and preprocessing

Clinical features are accessible for our 570 patients at TCGA. However, prior to processing, a lot of features are ignored due to the any of the following reasons:

- biological irrelevance for predicting cell senescence: corresponds to features which are uninformative to predict the G0-arrest label. This includes: name of the clinic in which the tissue was sourced, height, whether patient consent was verified, and number of first degree relatives with cancer diagnosis.

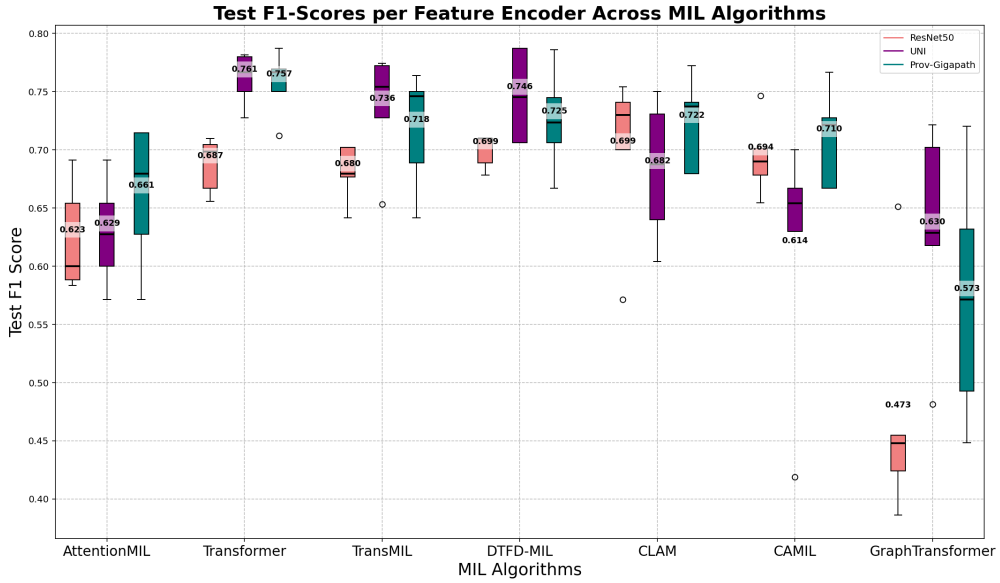

Figure 11: Boxplot of test F1 scores obtained from the 5 independent optimal models per fold, per feature encoder, evaluated over the test set. The often higher test scores (i.e. higher purple and teal bars) achieved by the UNI and Prov-Gigapath feature encoders suggests better generalization capabilities brought by foundation feature encoders in comparison to the standard ImageNet-pretrained ResNet50.

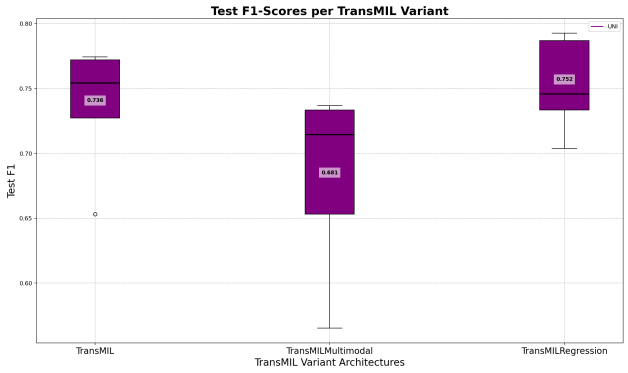

Figure 12: Boxplot of test F1 scores obtained from the 5 independent optimal TransMIL ablation models per fold, trained using the UNI encoder, evaluated over the test set.

- constant-valued variables: corresponds to features mostly filled with a constant value such as primary lymph node presentation assessment where 98% of the values were YES.
- semantically-same variables: corresponds to features which arguably refer to the same measurements, and thus were dropped to avoid multicollinearity. For example, if we include count of lymph nodes as part of our multimodal model, we dropped count of lymph nodes by H&E and by IHC. Similarly, we drop ICD-O-10 for ICD-O-3, and exclude anatomic neoplasm subdivision because of ICD-O-3 site.

After this, preprocessing occurs as follows:

1. We split the train-validation-test set for the clinical patient dataset, and take care in normalizing the continuous variables avoiding train-validation and train-test leakage. We save the features as tensors per patient for each CV fold and test set which is accessed separately during model training and evaluation.
2. A lot of variables concerning radiation therapy, e.g., drug administered, and its amount administered were dropped since they have a greater than 60% missing rate.
3. Variables like race and histological site have some of their values grouped to address class imbalance. For example, in our TCGA clinical dataset's training split, the variable 'race' consists of 4 values with ratios

Algorithm 1 Feature selection based on shadow features adapted from [Zukic, 2024].

Input: X_{train}, y_{train}, classifier, $n_{iter} = 100$, threshold$= 42$
Output: indexes of features selected from X_{train}
n, d = X_{train}.shape
scores = zeros(d) ▷ zero vector of shape d
X_{train} = join(X_{train},rand_col) ▷ join a random column of features to X_{train}
scale(X_{train}) ▷ min_max, normalize, robust_scaling, among others
for $i = 0 : n_{iter}$ **do**
    classifier(random_state = i).fit(X_{train},y_{train})
    feature_importances = get_feature_importances(classifier)
    rand_col_imp = feature_importances[-1] ▷ Get the random column feature's importance
    scores[argswhere(feature_ importances $>$ rand_col_imp)] ± 1 ▷ Count the times in which a feature's importance exceeds that of the random column feature's importance
end for
return argswhere(scores $>$ threshold)

indicating severe imbalance: White (76%), Black (20%), Asian (3%) and American Indian (1%). We thus group 'Black', 'Asian' and 'American Indian' under 'Non-White' and treat 'race' as a binary variable.

4. One variable per each one-hot encoded categorical variables is dropped to avoid multicollinearity. This is valid due to the mutual exclusivity of the values of the categorical variables. For example, one-hot encoding Pathological Stage with 9 possible values leads to the binary variables Pathological Stage I, Pathological Stage II(A, B), Pathological Stage III(B,C), and Pathological Stage IV(A) being formed. For example, a value of 1 for Pathological Stage IIA and 0 for the rest indicates this patient's CRC tissue is in Pathological Stage IIA. Since we assume cancer tissue cannot be at multiple stages simultaneously, and can only be in either of the described stages, Stage I is dropped to avoid collinearity as it is equivalent all remaining binary variables being set to 0.
5. One-hot encoding yields 30 features. We run a feature selection algorithm [Zukic, 2024] which selects 27 out of these 30 features. Feature selection (Algorithm 1) consists of training a classifier (in our case XGBoost) where a random feature vector is concatenated to the above preprocessed dataset to predict g0-arrest. Feature importances are computed, and for those with importance scores below that of the random feature vector's are recorded in a counter. Such process is repeated for $n = 100$ times, and we get rid of 3 features 'Pathologic Stage IIC', 'Pathologic Stage IIIA', and 'Pathologic Stage IVB' which for more than 42 times, their feature importances didn't exceed that of the random feature vector's.
6. This is finally followed by expert consultation with a computational biologist to ensure their relevance for multimodal fusion in our model.

We end up with the following list of clinical features:

- patient's age at the time of pathological diagnosis, which we treat as a normalized continuous variable.
- count of lymph nodes observable in the patient's tissue, which we treat as a normalized continuous variable.
- preoperative CEA level, which is treated as a normalized continuous variable. It refers to CEA in the blood before surgical intervention in CRC patients and serves as a tumor progression marker to guide therapy.
- gender, a binary variable with values 'male' and 'female'.
- race, a binary variable with values 'white' and 'non-white'
- other diagnoses, a binary variable indicating whether the patient has comorbidities
- pathological stage, a categorical variable with values stages II, IIA, IIB, III, IIIB, IIIC, IV, IVA. Stages II, IIA and IIB are also known as early stage cancer, while the remaining ones can be clustered under late stage cancer. Metastasis is one of the main markers differentiating these cancer stages.
- histological site, which is a categorical variable indicating tumor anatomical site following the Third Edition of the International Classification of Diseases for Oncology (ICD-O-3). Values include the 'cecum', 'colon, not otherwise specified (NOS)', 'rectosigmoid junction', 'rectum, NOS', 'sigmoid colon' and 'other'.
- patient's neoplasm cancer status, which is a binary variable indicating whether there's an observable tumor or not in the tissue.

- venous invasion, which is a binary variable referring to the presence of tumor cells within blood vessels outside the colorectal wall.
- lymphatic invasion, which is a binary variable referring to the presence of tumor cells within lymphatic vessel. Both venous and lympathic invasion are markers of metastasis and recurrence [Messenger et al., 2012].
- history of colon polyps, which is a binary variable indicating whether patient has developed polyps or not. Morphological details about the polyps are not provided.
- residual tumor, which is a binary variable indicating the presence of cancerous tissue after treatment, such as post-surgical resection.
- loss of expression of mismatch repair (MMR) proteins as detected by immunohistochemistry (IHC), which is a binary variable referring to whether there's a complete absence of nuclear staining for MMR proteins indicating inability to correct DNA replication errors. It serves as an biomarker for increased potential for tumorigenesis [Nadorvari et al., 2024].

D Hyperparameters of the MIL models benchmarked

We proceed in stating relevant hyperparameters of MIL models benchmarked.

	Epoch	**Initial learning rate, and weight decay**	**Optimizer**	**Learning rate scheduling policy**	**Additional hyperparameters**
AttentionMIL	32	$2 \times 10^{-5}, 1 \times 10^{-2}$	Adam	fit-one-cycle with a maximum learning rate of 1×10^{-4}, and the first 25% of the cycle with increasing learning rate (Wang et al., 2022)	
Transformer	8	$2 \times 10^{-5}, 2 \times 10^{-5}$	AdamW	cosine annealing decaying over training epochs with a minimum learning rate of 1×10^{-6}	
TransMIL	32	$2 \times 10^{-5}, 1 \times 10^{-2}$	AdamW	same as Transformer	
DTFD-MIL	42	$2 \times 10^{-5}, 1 \times 10^{-4}$	Adam for both tiers	learning rate decay starts at epoch 25 for both tiers by a factor of 0.2	5pseudo-bags
CLAM	42	$2 \times 10^{-4}, 1 \times 10^{-5}$	Adam	same as Transformer	dropout of 0.25 and 8 patches for instance-level clustering
CAMIL	30	$2 \times 10^{-5}, 2 \times 10^{-5}$	Adam	learning rate is reduced by a factor of 0.2 once a plateau in performance is identified	
GraphTransformer	42	$1 \times 10^{-3}, 5 \times 10^{-4}$	Adam	learning rate decay starts at epoch 20 by a factor of 0.1	
TransMILMultimodal	same as TransMIL				27 clinical features
TransMILRegression	same as TransMIL				MSE loss

Table 3: Hyperparameters adopted per MIL algorithm. For each algorithm, we embed the source where the hyperparameters are mentioned. We avoid hyperparameter tuning, and this includes not performing extensive neural architecture search. Unless stated otherwise, all models are trained by minimizing the BCEWithLogits loss. TransMILRegression is trained with the MSELoss.

E Interpretability analysis of Ensemble TransMIL with UNI feature encoder

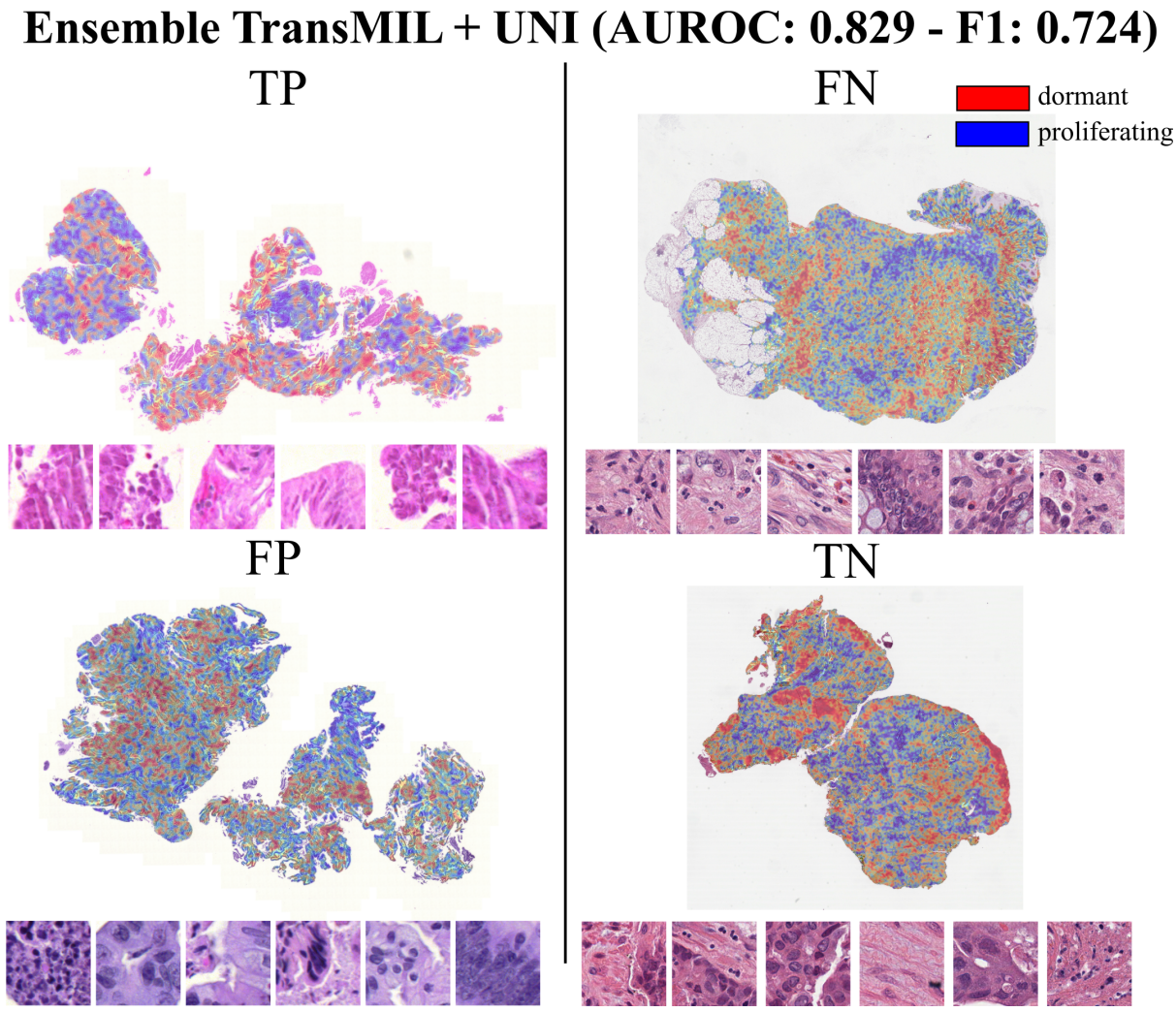

Figure 13: Heatmaps generated by the Ensemble TransMIL with the UNI feature encoder. We provide correct and incorrect classifications, and below each heatmap we append a sample of 6 patches according to their attention scores contributing to the slide-level prediction. For TP and FP, these patches have the highest attention scores explaining a positive prediction. For TN and FN, the patches have the lowest attention scores explaining a negative prediction.

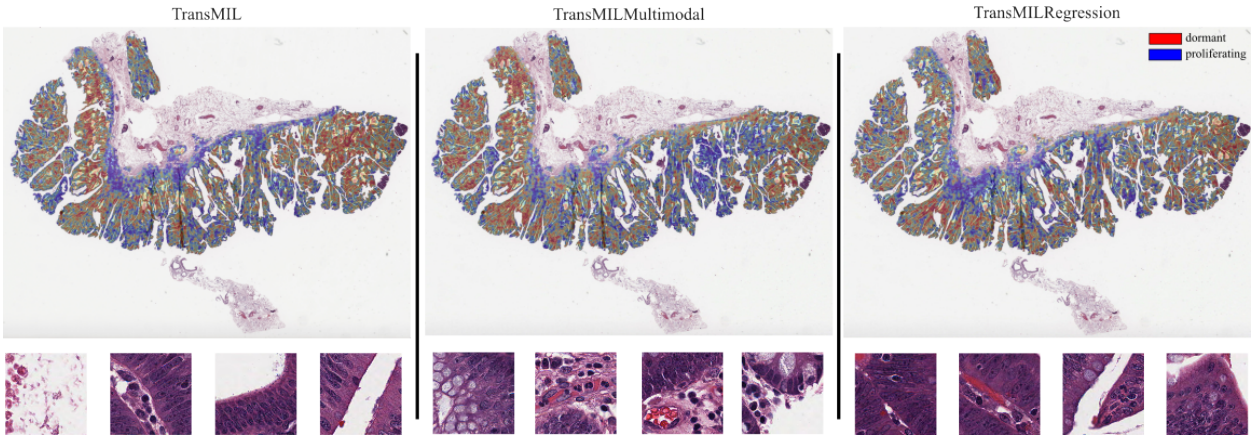

Figure 14: Side-by-side comparison of heatmaps generated by ablations of TransMIL with the UNI feature encoder. Below each heatmap is a sample of 4 patches with the highest attention scores contributing to the prediction of G0-arrest, and are all TP predictions. For TransMIL and TransMILMultimodal, this corresponds to a prediction of 1, while for TransMILRegression, this is a negative score of -0.39 with ground truth -2.1 binarized at ≤ 0