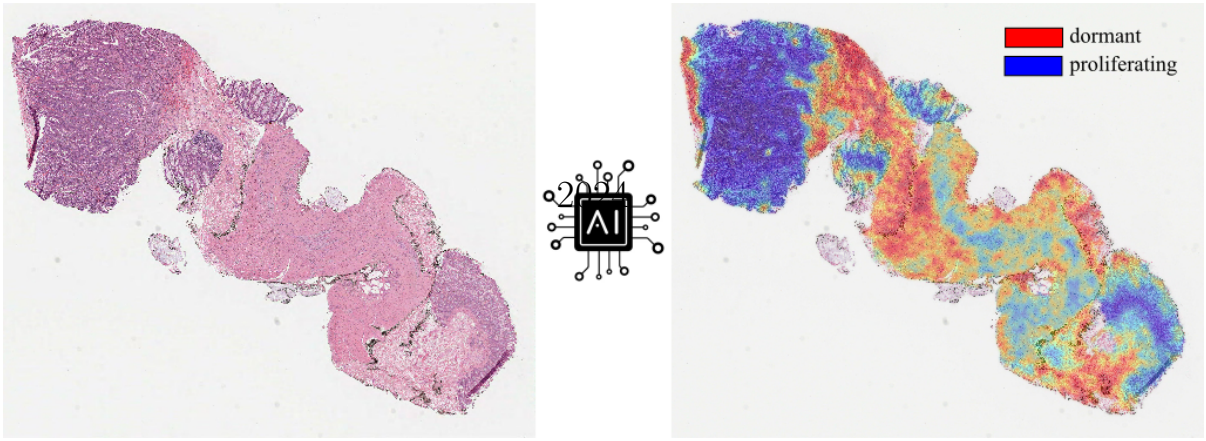
LONDON'S GLOBAL UNIVERSITY
Department of Computer Science

Unveiling Colon Cancer Therapy Resistance: A Multiple-Instance Learning Approach to Cell Dormancy Classification in Histopathology Images

An Xuelong

Supervisors: Dr. Pan Shi, Dr. Maria Secrier, Dr. Petru Manescu



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.

Keywords— G0-arrest - Colon cancer - Multiple instance learning

Acknowledgements

I would like to express my gratitude to Dr. Maria Secrier for giving the opportunity to work on this project where I learnt a lot about the intersection of cancer biology and deep learning. I want to especially thank Dr. Pan Shi for his guidance on experiment design, for writing HistoMIL which served as the backbone of my thesis, and for sharing with me the latest trends in deep learning research which greatly helped me grow as an applied researcher. I also want to thank Dr. Petru Manescu for his supervision and feedback on my thesis.

Doing a masters in London has been a challenging journey, and I want to greatly thank the friends I made here for their support which gives the strength to overcome the different difficulties I've faced, as well as the unforgettable memories I've made. I want to thank Mediha for your friendship, with whom I extend my gratitude academically also since I borrowed her feature selection algorithm for my thesis. I want to thank Kenza for organizing WednesdAIs, weekly group discussions where I got priviledge to share with each other our ongoing work and offer feedback to each other to help each other improve. I also want to thank Luba for providing feedback to my work, and helping me regarding my post-master career plans. I want to thank Tariq, Adam, Sheng Hui, Amy, Josie and Isha for making my London experience fun, with activities ranging from podlucks, travelling in Cambridge, exploring Central London, going to the gym and going to festivals together. May our fates intertwine and see each other again in the near future.

I want to finally express my gratitude to my family for their love and support throughout my life.

"*... [J]ust as mathematics turned out to be the right description language for physics, AI may turn out to play a similar role for biology.*"

— Demis Hassabis

Table of Contents

List of Definitions, Abbreviations and Synonyms

We begin by defining the following list of terms which consists of widely used acronyms, and terms which have been called with multiple names but share the underlying meaning.

AI: Artificial intelligence

AUROC: Area Under the Receiver Operating Characteristic Curve. In the literature, authors also abbreviate it as ROC-AUC, OR AUC

CEA: Carcino-Embryonic Antigen

CNN: Convolutional neural network

CRC: colorectal cancer

CV: Cross-validation

DL: deep learning

DNA, mRNA, miRNA: Deoxyribonucleic Acid, messenger ribonucleic acid, micro-RNA

G0-arrest cells: tumor cells in G0-arrest are thought to be cancer therapy-resistant. In our work, we interchangeably describe them as 'persister', 'dormant', 'arrested', 'quiescent' (for G0-arrest cells that can revert back to the cell cycle), 'senescent' (for cells who entered into an irreversible G0-arrest stage) [Santos-de Frutos and Djouder, 2021]. We adopt the mutual exclusivity assumption, which entails that cells not in G0-arrest are undergoing the cell-cycle, and we interchangeable describe these as 'normal-cycling', or 'proliferating'.

GPU: Graphics Processing Unit

H&E: Hematoxylin and eosin

HRD: Homologous Recombination Deficiency

IHC: Immunohistochemistry

IID: Identically, Independently Distributed assumption in statistics

MIL: Multiple Instance Learning

MSI: Microsatellite Instability

PCC: Pearson's Correlation Coefficient

Patches: regions of a WSI cropped to process them in parallel. They're also called tiles, or instances in MIL

RFC: Random Forest Classifier

ROI: Regions of Interest

ST: Spatial transcriptomics

SoTA: State of The Art

TCGA, YCR-BCIP and CPTAC: The Cancer Genome Atlas, Yorkshire Cancer Research Bowel Cancer Improvement Programme, Clinical Proteomic Tumor Analysis Consortium. Each is a cohort with public cancer data.

WSI: whole slide image. They're also called slides, or bag in MIL

scRNA-seq: single cell ribonucleic acid-sequencing

Common Notation

- x_d^n: measurement of d-th feature of n-th datapoint, e.g., pixel value
- $\mathbf{x}$: in bold we denote a vector of random variables
- $(X, \mathbf{y}) \in \mathcal{D}$: Dataset consisting of input features and labels
- $\mathcal{L}(\mathcal{D}|\theta)$: Likelihood function of dataset given parameters θ
- $-\mathcal{LL}(\mathcal{D}|\theta)$: Negative log likelihood function of dataset given parameters θ
- $\mathcal{W} \in \mathbb{R}^{N \times D}$: a feature representation for a single whole slide image of dimension $N \times D$ where N is number of patches and D is number of features per patch.
- $\mathcal{A} \in \mathbb{R}^{N \times N}$: an adjacency matrix representation for a single whole slide image of dimension $N \times N$ where N is number of patches.
- $\mathbf{f} \in \mathbb{R}^{27}$: clinical feature vector for a patient.

1 | Introduction and Background

1.1 Colorectal Cancer Incidence

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. Several modifiable lifestyle factors contribute to the development of CRC, such as a high intake of processed meats and low intake of fruits and vegetables, sedentary lifestyle, obesity, smoking, and excessive alcohol consumption [Organization, 2023]. As per $2016-2018$ statistics, bowel cancer is the 4th most common cancer in the UK, accounting for 11% of all new cancer cases. In females in the UK, bowel cancer is the 3rd most common cancer (10% of all new female cancer cases). In males, it is the third most common cancer (12% of all new male cancer cases) [UK, 2015].

A large proportion of CRC incidence and mortality is preventable through regular screening, surveillance, and timely high-quality treatment [Siegel et al., 2023]. 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., 2024a], 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]. The drivers of recurrence are an ongoing research area.

1.2 G0-Arrest and Cancer Recurrence

'Dormant' or 'persistent' cells have been garnering the attention of the research community for their role in relapse. Typically, an eukaryotic cell cycle can be split into four phases based on the timing of DNA synthesis: 1) a G1 phase (gap 1), which corresponds

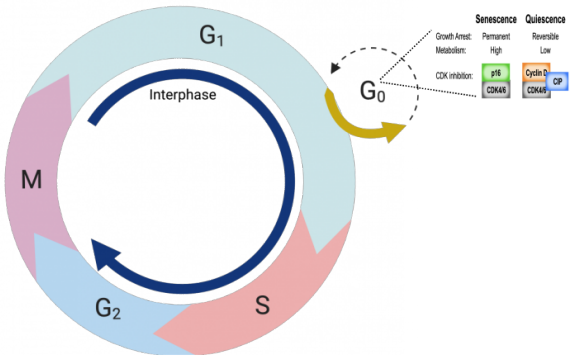


Figure 1.1: A complex interplay of signals drive a cell's entry into G0-arrest or exit thereof. Crucial to entry to G0-arrest is the inhibition of cyclin-dependent kinases [Pack et al., 2019]. Figure is obtained by editing from the above sources and [Hardy,].

to the interval between mitosis and initiation of DNA replication, then 2) an S phase (synthesis), during which DNA replication takes place, followed by 3) the G2 phase (gap 2), during which cell growth continues and proteins are synthesized in preparation for 4) mitosis (M). Cells can exit this replicative cycle into a state of **'G0-arrest'**, in which although they might be metabolically active, they cease to grow and have reduced rates of protein synthesis [Cooper, 2000]. These cells are 'quiescent' if they can revert back to the cell cycle from G0-arrest, otherwise 'senescent' (Figure 1.1).

During carcinogenesis, the role of 'dormant' cells has been studied along metastasis and relapse [Gao et al., 2017]. For instance, cells in G0-arrest are resistant to anti-cancer compounds, such as chemotherapy, that target actively dividing cells. Furthermore, G0-arrest cells also 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 [Wiecek et al., 2023]. Cell dormancy can be caused by a variety of factors, whether it's induced through environmental stress as shown by simulations from in-vitro models [Mitra et al., 2018], replicative stress, oncogene ac-

tivation, or could be a natural stage of a cell's developmental process [Oki et al., 2014].

Given their relevance for predicting relapse, [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, it comes with broad clinical implications involving the monitoring of this state in a tumor through sequencing technologies to study therapeutic resistance.

Indeed, the integration of bulk, single-cell, and spatially resolved sequencing techniques provide unprecedented insights into characterizing the TME of CRC, furthering our understanding of how cancers grow and spread. For example, [Joanito et al., 2022] use bulk transcriptome sequencing and single-cell RNA sequencing (scRNA-seq) to identify colorectal cancer molecular subtypes (CMS) and associate the mesenchymal CMS4 subtype with poor relapse-free survival. Complementing these sequencing technologies are more powerful, spatially-resolved transcriptomics (ST), which provide insights into the spatial organization and interactions between tumor and stromal cells within the CRC tissue, helping reconstruct the putative interaction networks between tumor cells and their microenvironment, identifying potential therapeutic targets and biomarkers for determining CRC molecular subtypes [Xiao et al., 2024b]. For instance, research has shown that a spatial genetic signature can discriminate neoplastic from non-neoplastic compartments in colon cancer, serving as biomarkers for relapse [Sallinger et al., 2023]. These findings emphasize the importance of characterizing tumor heterogeneity in a spatial context using next-generation sequencing, in which by using the G0-arrest signature helps gauge dormancy and inform the trajectory of tumor growth.

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

[Levy-Jurgenson et al., 2020]. Furthermore, sequencing technologies, especially spatially-resolved ones, may face several hurdles when translated into routinely available, clinical-use due to their novelty, main usage in research/experimental settings, associated costs [Smith et al., 2024] 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.

1.3 Deep learning, Whole Slide Imaging and Molecular Profiling

Stained human tissue is the gold standard for the assessment of many diseases including cancer. The most common stain is H&E which highlights the nuclei (stained with hematoxylin) and cytoplasmic/extracellular components (stained with eosin) of tissue samples on glass slides. It is applied in nearly all clinical cases, covering 80% of all the human tissue staining performed globally [de Haan et al., 2021]. The digitization of H&E stained-cancerous slides into high-resolution whole slide images (WSIs) has emerged as a vital resource for prognosis prediction as they richly capture the visual details of tissue structure and cellular morphology, which can be used to analyze the TME landscape [Lee, 2023].

Associated with the availability of such wealth of data is the advent of deep learning (DL) algorithms. DL has revolutionized numerous fields, particularly within the medical sciences. In oncology, 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, biomarker localization, (pan-)cancer subtype classification, and prognosis prediction [Song et al., 2023,

Tran et al., 2021, Couture, 2022, Lee, 2023]. In particular for colon cancer, convolutional neural networks (CNNs) have been trained to detect it from WSI [Alboaneen et al., 2023], classify tumor-immune cells from colon tissue [Parreno-Centeno et al., 2022], distinguishing between microsatellite instability (MSI) and microsatellite stable (MSS) subtypes in colorectal WSI [Hezi et al., 2024], classifying homologous recombination deficiency (HRD) and MSI spots directly from CRC WSI [Schirris et al., 2022], detect multiple genetic mutations [Konishi et al., 2023], among others. In this regard, we ask how can we exploit DL technologies for assessing relapse given solely an WSI of H&E -stained tissue, which could come with benefits of reducing costs of sequencing and improving accessibility and accuracy?

While at first glance it is an ill-posed problem, the powerful feature extraction capabilities of DL-based models can extract relevant tissue structural information from the colon's H&E WSI to classify a dormant cell, especially given the causal link between genes and phenotype. While experienced pathologists can identify a few putative biomarkers in H&E stained tissue, inter-and intra observer variability can arise when assessing a sample. Instead, we could benefit from the usage of DL-based tools given their prospects of improving the accuracy of and speeding up the screening of H&E tissue [Parreno-Centeno et al., 2022]. This would assist pathologists in rapidly stratifying patients based on their dormant cell population.

Furthermore, literature evidence has identified that cells undergo morphological changes throughout the cell cycle [Dapena et al., 2015]. In the context of in vitro models of cellular dormancy in primary fibroblasts and other types of cells, biomarkers have been identified for detecting dormancy, mainly revolving around detecting changes in gene expression patterns since cells at this stage cease dividing [Mitra et al., 2018]. Dormant cells have also been characterized by morphological alterations, including larger, flat bodies and organelle abnormalities, as well as loss of physiological functions due to their inability to proliferate [Huang et al., 2022]. G0-phase is also characterised by low metabolic

activity, along with a decrease in the production of ribosomal RNA and proteins, leading to reduction of their volume and size [Santos-de Frutos and Djouder, 2021]. Thus, because G0-arrest cells may exhibit changes that reflect their non-proliferative state, such as altered cell size and nuclear morphology (although these changes can be subtle and context-dependent), we hypothesize this is enough to pave the way for DL models equipped with their deep feature extraction capabilities to identify these traits in WSI.

Given this premise, in this work, we define the following aims and objectives:

1. Thoroughly benchmark different DL architectures to predict G0-arrest in colorectal WSIs. We perform ablation studies to study which architectural components contribute to high classification accuracy.

2. Perform interpretability analysis of the models to understand their output decisions and discuss their biological significance.

3. Contribute to the computational histopathology community a comprehensive DL pipeline for analysing WSIs for biomarker prediction.

2 | Literature Review

Computational histopathology is a multiplexed field where much work has been done. In this section, we non-exhaustively summarize studies relevant to the design of our DL pipeline and prediction task.

2.1 Challenges in Working with Whole Slide Images

The gigapixel resolution and thus complexity of WSIs present unique computational challenges for the design of a DL pipeline to analyze them. Broadly, such DL pipeline can be split into three phases: *feature extraction*, *training/validation* and *inference/interpretability analysis*.

2.1.1 Feature extraction

The typical paradigm of pre-processing WSIs consists of 1) segmenting the tissue whilst excluding any holes from the background, followed by a patch-wise cropping method which divides the gigapixel tissue into thousands of image patches with smaller dimensions, e.g., 224×224 pixels (a process known as 'tessellation'). This is because gigapixel WSIs cannot be processed as a whole using modern deep CNNs, let alone transformer-based neural networks, mainly due to limited GPU memory [Gadermayr and Tschuchnig, 2024]. These image patches can either be passed directly to a model for making a prediction, or fed into 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 final global prediction results [Tan et al., 2023].

2.1.2 Training and Validation

The classical paradigm of training DL models is through fully-supervised learning, whereby each datapoint $(\mathbf{x}, y) \in \mathcal{D}$ is annotated with a relevant label y. In the context of tessellated WSIs, this would entail each image patch, or in extreme cases, each pixel, needs a corresponding diagnostic label such as the presence/absence of senescent cells. Such intricately labeled data could be available such as in [Chen et al., 2020] where they train a CNN for tumor grading in a WSI through supervised learning, or in ST datasets where each 'spot' in a WSI is matched with scRNA-seq data. However, it is often the case that only slide-level labels are available due to the intense annotation burden associated with WSIs [Tan et al., 2023, Gadermayr and Tschuchnig, 2024], as well as the expensive costs of using ST platforms. This leads to non-spatially resolved bulk-RNA or scRNA-seq data to be more readily available.

A slide-level label only makes a broad statement about the WSI, i.e., if only in certain regions of the tissue G0-arrest cells are identified, then the entire WSI receives a positive label. The common practice to train a DL model in such a setting is to resort to a weakly-supervised learning framework, in particular ***multiple-instance learning***. In the literature, a slide is referred to as a 'bag', and a patch is referred to as an 'instance'. The goal of training a model under the MIL framework is to learn to classify slides, as well as the key patches that 'trigger' the slide's label. Interestingly, learning such key patches, $p(\text{label}|\mathbf{patches})$, enables the DL model to highlight regions of interest (ROIs) in the slide as part of interpretability analysis [Ilse et al., 2018].

Mathematically, a bag B^n is a collection of instances $\{x_1^n, x_2^n, ..., x_d^n\}$, where each B^n is given single label y^n as follows:

$$y^n = \begin{cases} s & \text{if } \exists j \text{ such that } x_j^n = 1 \\ 0 & \text{if } \forall j, x_j^n = 0 \end{cases} \tag{2.1}$$

, 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.

We train a classifier f that can predict the label of a new bag based on patch embeddings, which involves optimizing a cost function (the negative log likelihood) of the parameters θ:

$$-\mathcal{LL}(\mathcal{D}|\theta) = \sum_{i=1}^{N} \ell(y^i, \hat{y}^i) \tag{2.2}$$

, 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.

Equation 2.2 is optimized with respect to θ via gradient descent routines for which there is a plethora of libraries implementing them. 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.

2.1.3 Inference and Interpretability Analysis

Evaluation is done at a WSI-level due to the lack of patch-level annotations. In the context of binary classification, a model's predictive performance is commonly measured via the Area Under the Receiver Operating Characteristic (AUROC) of the model's

predictions over unseen test WSIs. The AUROC measures the tradeoff between the true positive (TP) rate (sensitivity) and false positive (FP) rate (specificity) at different classification thresholds. An AUROC closer to 1 indicates a high-performant classifier, while a value of 0.5 represents a random classifier. Because in cancer histopathology, benign tissue samples often heavily outweigh malignant samples [El Nahhas et al., 2024, Gadermayr and Tschuchnig, 2024], the F1-Score= $2 \times \frac{precision \times recall}{precision + recall}$ is also used to account for class imbalance. It represents the harmonic mean of precision[1] and recall[2], where closer to 1 indicates better classification performance. This is because in a highly imbalanced dataset, alternative metrics like precision can be overestimated simply because the trained f is biased towards identifying benign samples.

Additionally, interpretability analysis can be carried out by using an MIL algorithm's patch-level predictions [Gadermayr and Tschuchnig, 2024, Campanella et al., 2019] prior to pooling. These local predictions can be mapped over the original WSI input space to visualize which regions contribute most to the prediction. Only in some cases, such as with well-established MIL histopathological benchmarks like CAMELYON16, pathologists have exhaustively annotated ROIs relevant to the task at hand, such as for the presence of metastases [Khened et al., 2021]. Then, evaluation at a qualitative level can be done comparing the model's local predictions with the ground-truth annotation. In our case, however, we lack pathologists annotating regions of proliferating and dormant cells, thus we can't resort to such localized evaluation. Alternative evaluation methods such as in [Parreno-Centeno et al., 2022] resort to ST, which uses the gene expression data spatially resolved over the WSI tissue, one can compute at a 'spot' level a label of interest, such as cell dormancy at that 'spot'. Then, a model's patch-level predictions can be compared with this 'spot'-level ground-truth. However, this approach is more nuanced due to the patch size being inconsistent with the 'spot' size. Additionally, ST data is not necessarily available for the same datapoints we use and is more expensive to

[1] ratio of true positives to true and false positives (FP)
[2] ratio of true positives to true positives and false negatives (FN)

obtain.

2.2 Related work

We proceed in describing prior work which inspires the design of our pipeline and models:

Deep learning for molecular-level prediction: Computational histopathology initially explored DL for predictive tasks at a histological level, such as tumor staging or cancer sub-typing, to aid patient stratification. Over the years, the field has matured to train models for molecular-level predictions given solely the WSI, including MSI, HRD, genetic mutations like BRAF, mRNA and miRNA expression, among others [Couture, 2022]. In contrast to much prior work focusing on biomarker classification, [El Nahhas et al., 2024] propose a model which predicts biomarker scores rather than categorical labels of cells in H&E images. They argue 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 for predicting HRD labels and scores, they found that using 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 AUROCs 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 in this work.

Foundational models for histopathology: There is an increased interest in the training and publication of foundational models thanks to the massive size and diversity of the training data that is available for representation learning. One of the benefits of foundational models is that their feature extraction capabilities far surpass that of pretrained convolutional-based networks, such as ResNet, which has been usually used as the go-to feature extractors for WSI patches. ResNet-based architectures which are pretrained on natural images (and sometimes finetuned to histopathology datasets) are seldom powerful feature extractors which served as the backbone for much work achieving SoTA in computational histopathology [Tan et al., 2023]. Pretrained feature extractors that are localized to histopathology such as CTransPath [Wang et al., 2022] and REMEDIS [Azizi et al., 2023] are already powerful feature extractors since they are pretrained through self supervised learning on pan-cancer tissue types sourced from TCGA. Foundation models go a step beyond by scaling both the size of the architecture being trained and the datataset being used. For instance, while CTransPath and REMEDIS are mainly pretrained on TCGA (around 20000 diagnostic slides), consisting mainly of primary tumor sites, foundational models are trained on above 100000 WSIs spanning patient cohorts, cancer tissue types, and across diagnostic tasks. Thanks to such diversity, features output by foundational models are context-dependent, and semantically rich. These can help alleviate the demand for high volumes of data for representation learning, which is relevant for our purposes where we only have 578 colon WSIs available for training and test, as well as reflects the general problem of scarce annotations in computational histopathology. These features also have high prospects of generalization given their pretrained regime across tissue types [Chen et al., 2024]. Indeed, generalizability is a major challenge in computational histopathology, as it is defined as a model's capability to achieve high prediction performance on unseen WSIs which exhibit variability with respect to the training data in terms of image acquisition protocol, medical center,

and inter-personal phenotypic differences. Therefore, we can ask whether foundational feature encoders enables MIL models to generalize.

In computational histopathology, foundational models like Virchow [Vorontsov et al., 2023] can achieve state-of-the art performance in several downstream prediction tasks concerning biomarker prediction and tumor subclassification with no training (zero-shot learning) or limited finetuning. [Chen et al., 2024] propose UNI, a foundational 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. By accounting restricted computational resources, we are faced with a few limitations: first, we are unable to fine-tune these foundational models for representation learning given their computational architecture which demands high-end GPUs. As such, we only restrict ourselves in using them as frozen feature extractors that provide semantically rich embeddings used by our main model within MIL.

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 in-

teractions with the hope of improving the model's predictive expressivity and accuracy [Feng et al., 2024]. There exist many fusion methods in supervised learning, ranging from a simple concatenation of different input modalities (early fusion), intermediate fusion, or fusion of the embeddings of the different input modalities before making a decision (late fusion) [Stahlschmidt et al., 2022]. In computational histopathology, [Chen et al., 2022] propose a pan-cancer model integrating WSI with genomic data through late fusion 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. [Volinsky-Fremond et al., 2024] also combines through late fusion the tumor stage with endometrial H&E WSI 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 in our multimodal model 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].

Spatial inductive biases: Most MIL algorithms used over the years assume permutation invariance of the image patches [Ilse et al., 2018], thus models trained through MIL don't necessarily capture the spatial dependency of image patches. This spatial dependency can be accounted through different approaches, such as imposing a geometrical inductive bias in the model through graph neural networks (GNNs). [Yacob et al., 2023] propose using a graph transformer to capture the spatial information of patches within a WSI, and they train it to detect subtypes of basal cell carcinoma (BCC) through MIL, achieving test accuracies of 93.5%, 86.4%, and 72.0% for binary, ternary, and 5-class BCC subtype classification tasks. [Eastwood et al., 2023] propose a CNN-based GNN for classifying 3 subtypes of mesothelioma, and achieve an AUROC of 0.86 on an ex-

ternal validation set. However, neither author performs an ablation study analyzing the benefits of using the GNN compared to deep neural networks, or transformer-based networks. Thus, whether employing GNN actually help in improving classification accuracy over alternative DL architectures within a MIL setting in processing WSI is still an open question, and we leave the use of GNNs for future exploration[3].

Spatial constraints can also be imposed over the input space by representing the WSI as a graph, where patches are nodes, and the MIL model only focuses on patches adjacent to one another. This form of contextual constraint helps consider dependencies amongst patches, alleviating the permutation invariance assumption above [Zheng et al., 2022, Fourkioti et al., 2024]. In our work, we explore the impact of these spatial constraints over the input space in G0-arrest prediction.

Interpretability analyses: DL models are known to be black-box predictors, and as such much effort is poured into making their predictions justifiable, interpretable to the human eyes, and hence trustworthy when assisting clinical pathologists. There exists a plethora of interpretability methods tailored to computational histopathology. [Lu et al., 2021] propose mapping the attention scores associated to each patch back to the original WSI to plot a heatmap 'explaining' a slide-level prediction. Such heatmap is color-coded based on the task at hand, where for binary classification we can interpret patches with high attention scores justify the prediction of class 1, while patches with low attention scores justifying a slide-level prediction of 0. Alternative approaches like LIME, SHAP and/or GradCAM wouldn't be appropriate in our problem setting since they require us to perturb the original input space. Nonetheless, recall that due to their gigapixel size, we work with a feature representation $\mathcal{W} \in \mathbb{R}^{N \times D}$ of a WSI, instead of the WSI itself.

[3]To use GNNs, we need further preprocessing, as each WSI's feature representation must adopt the semantics of a graph, where feature vectors are nodes. Furthermore, they have to be compatible with graph dataloaders, while our current HistoMIL pipeline implements a standard PyTorch dataloader.

2.3 HistoMIL

MIL packages: 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 3.1, 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.

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 2.1). 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 protocol at [Lu et al., 2021]. This involves tissue segmentation consisting of reading the WSI at a downsampled resolution, then converting it from RGB (red-gree-blue) to HSV (hue-saturation-value) color space. A binary mask for the tissue regions (foreground) is computed based on thresholding the saturation channel of the WSI after median blurring to smooth the edges, which is followed by additional morphological closing to exclude small gaps and holes (Figure 3.1.1.b). The segmented tissue is tessellated with a patch size of (224×224) with no overlap (Figure 3.1.1.c). 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 ($D = 1536$), where UNI and Prov-Gigapath are foundational feature encoders (Figure 3.1.1.e). 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)

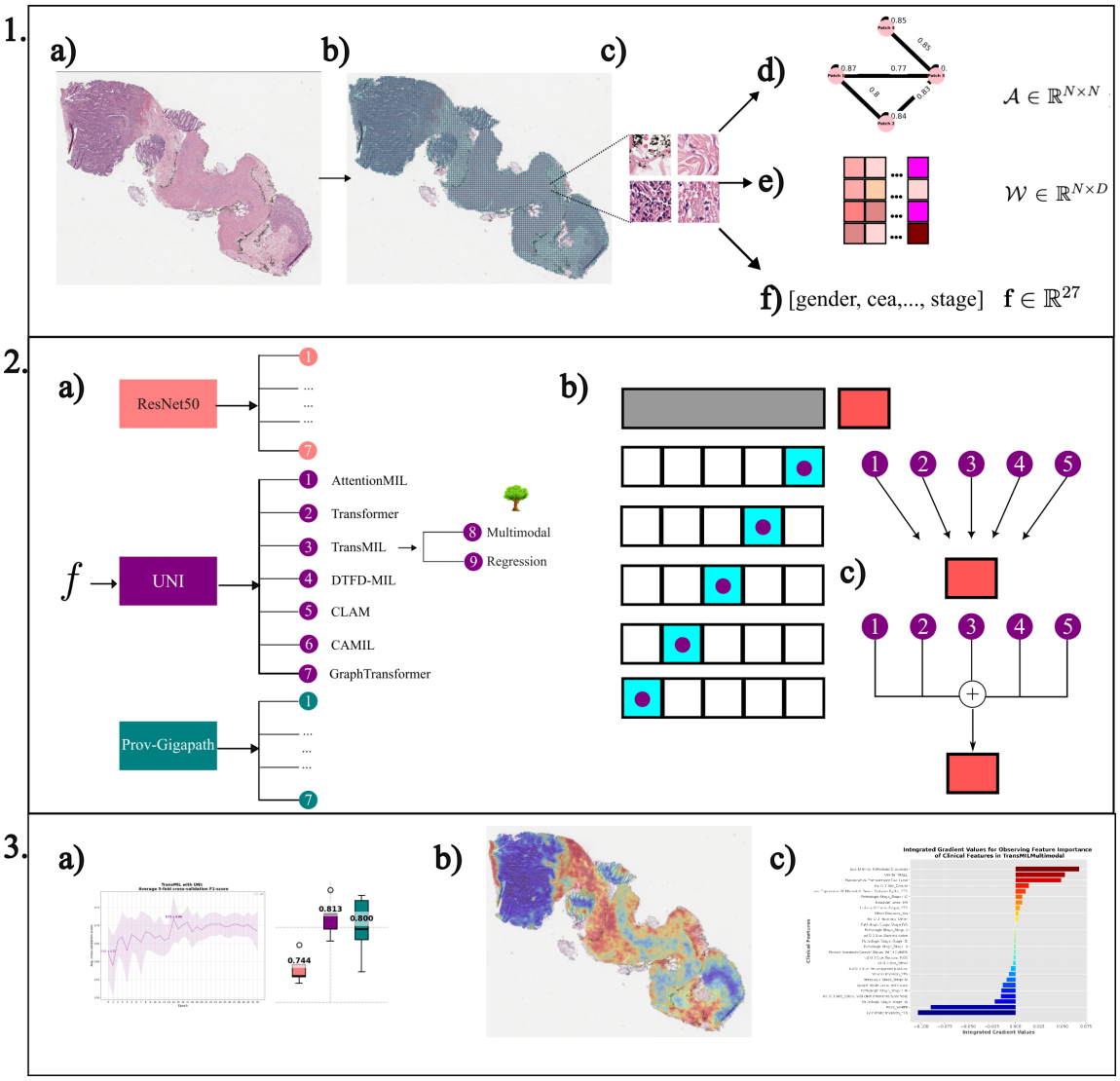


Figure 3.1: Depiction of our pipeline. First, 1.a) each colorectal WSI 1.b) undergoes tissue segmentation and 1.c) tessellation. At 1.d), we compute an adjacency matrix $\mathcal{A}$ via Equation 3.1. For each WSI, 1.e) we compute $\mathcal{W} \in \mathbb{R}^{N \times D}$ using 3 feature encoders: ResNet50, 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 3.2. 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.

it can range between $[10000, 90000]$. In the interest of training some MIL algorithms with topological constraints, a tessellated 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 (Figure 3.1.1.d). 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 3.1:

$$a_{ij} = \begin{cases} \exp\left(-(\mathbf{h}_i - \mathbf{h}_j)^2\right) \text{ iff } (v_i, v_j) \in E, (\mathbf{h}_i, \mathbf{h}_j) \in \mathcal{W} \\ 0, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{3.1}$$

, 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 according to their embeddings [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 (Figure 3.1.1.f):

- 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 carcinoembryonic antigen (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].

We discuss in detail the selection and preprocessing of the above features at the Appendix 6.1, 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. Exploratory data analysis of our clinical features can be visualized at https://github.com/awxlong/scripts_g0_arrest/blob/main/step_1_preprocessing/clinical_features/sweetviz_report_selected_features.html.

3.2 Benchmarking models under 5-fold CV

We are faced with the classical challenge of data shortage in histopathology, where we only have 578 slides. To ensure our benchmarked models learn the necessary morphological features displayed over the WSI to discriminate G0-arrest cells, we perform a $90\% - 10\%$ train-test split, yielding 58 images for test-evaluation. To guarantee model robustness, we accompany this with 5-fold cross validation (CV) in the training set, which consists of splitting the training set into 5 non-overlapping folds, where 4 are used for training and 1 for validation a model. This is repeated 5 times, each time the model is validated on 1 different validation fold, and trained on the remaining 4 folds (see Figure 3.2).

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, which we discuss below.

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. By ease of adoption, we avoid algorithms such as Distillation Across Scales-MIL (DAS-MIL) [Bontempo et al., 2023] as they require different features matrices per WSI corresponding to different magnifications of the slide, while all algorithms we benchmark only require $\mathcal{W}$ at the slide's original resolution. The MIL algorithms, along with a brief justification, are:

- **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 con-

sists 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]. The fusion of clinical features enables us to explore 2 interesting ideas. On one hand, we can confirm whether clinical features alone suffice for predicting G0-arrest, which allow us to underscore (or not) the advantage of training DL models looking at morphological features instead. We do this by comparing the above results with an ensemble of random forest classifiers (RFC) trained solely on clinical features. Its training follows the same regime as all the MIL algorithms above (Figure 3.2): 5-fold CV, followed by an ensemble on the test set. Secondly, the conditioning of the G0-arrest decision on the joint clinical features and their morphological context allows us to inquire through interpretability analysis which clinical parameters gain prominence in TransMILMultimodal. We can then study which clinical parameters are important for predicting G0-arrest it it's not conditioned jointly with the morphological context. This prompts us to ask, if TransMILMultimodal focuses on different clinical features than the RFC, why is that the case?

– 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 2.1)[1].

[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.

- **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 3.1 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,

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

can complete their 5-fold CV in $\leq$ 16 GB of GPU memory in less than 3 days. The Transformer is the only MIL algorithm which requires an A40, corresponding to 48 GB of GPU memory, and can complete the 5-fold CV regime in less than 6 hours. We also don't perform hyperparameter tuning due to constrained computational resources, and instead reuse the hyperparameters mentioned in their respective papers. For specific details, please see Appendix 6.2.

AUROC is the de-facto metric for assessing deep learning model classification performance in computational histopathology, as well as it's appropriate for making cross-modal comparisons because it's agnostic to decision thresholds. We accompany it by the F1-score [Schirris et al., 2022] by binarizing class probabilities at a sound threshold of 0.5 given our balanced distribution of G0-arrest binary labels. Furthermore, F1 is arguably a more valuable metric for our purposes as it penalizes a model's FP and FN predictions, which is important if the predictions have implications on elucidating how a tumor spreads. In particular for TransMILRegression, where outputs stop being probabilities, we compute instead the Pearson's correlation coefficient (PCC) between them and the ground-truth G0-arrest scores [El Nahhas et al., 2024]. Because we can still binarize them at a clinical threshold of 0, we can compare F1 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

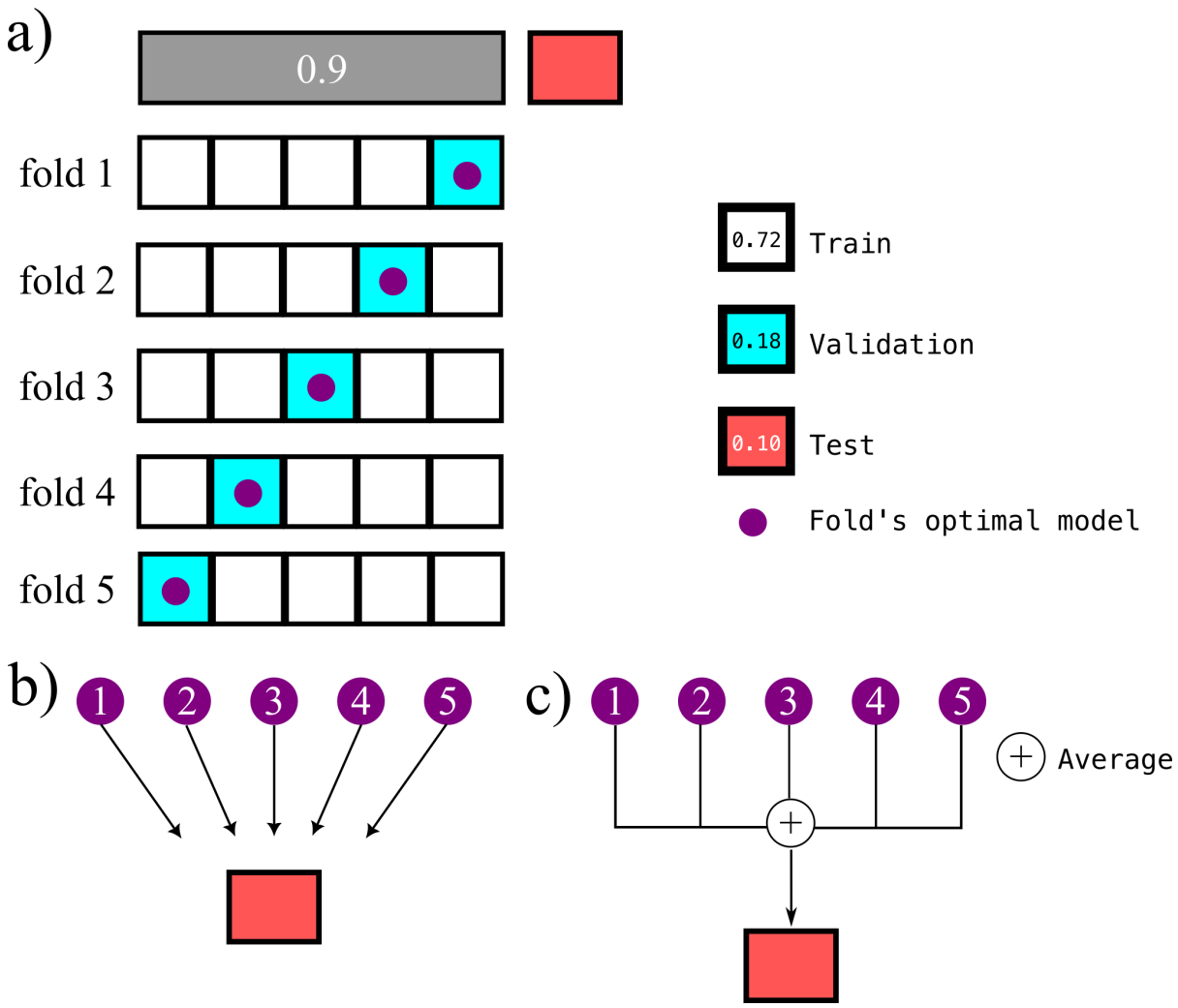


Figure 3.2: Illustration of our training, CV and evaluation framework with ensembles. a) illustrates train-test splitting, followed by 5-fold CV where our training set is split into 5 equally-sized folds, where 4 are used for training and 1 for validation. Per fold, HistoMIL checkpoints the optimal model 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 per fold, 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.

fusion and regression. This explains why only TransMILMultimodal and TransMILRegression are amongst the benchmarked models above (Figure 3.2a).

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 3.2b).

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 3.2c).

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]. Because in all MIL algorithms, the attention score per patch is pooled to make a prediction (see explanation of Equation 2.2), 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 pre-

diction 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. For the random forest binary classifier, we can look at relative feature importances to understand which features contribute the most to the final prediction. Each relative feature importance is a normalized value which aggregates the reduction in entropy achieved by each feature across all trees in the forest, where higher feature importance indicates a greater contribution to the model's predictions.

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

4 | Results and Discussion

We proceed in discussing the results of our extensive experiments. We show both computational and biological insights regarding how DL *can* aid clinical pathologists in gauging the G0-arrest population in colon tissue.

4.1 Foundational model-based feature encoders can help with generalization

We benchmark the MIL models and evaluate them on their predictive accuracy on G0-arrest. Our cross-validation (CV) results for each feature encoder and MIL algorithm are shown at Figure 4.1 for the AUROC metric and at Figure 4.2 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 foundational feature encoders didn't show much performance improvement.

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 4.3 because PCC is not available

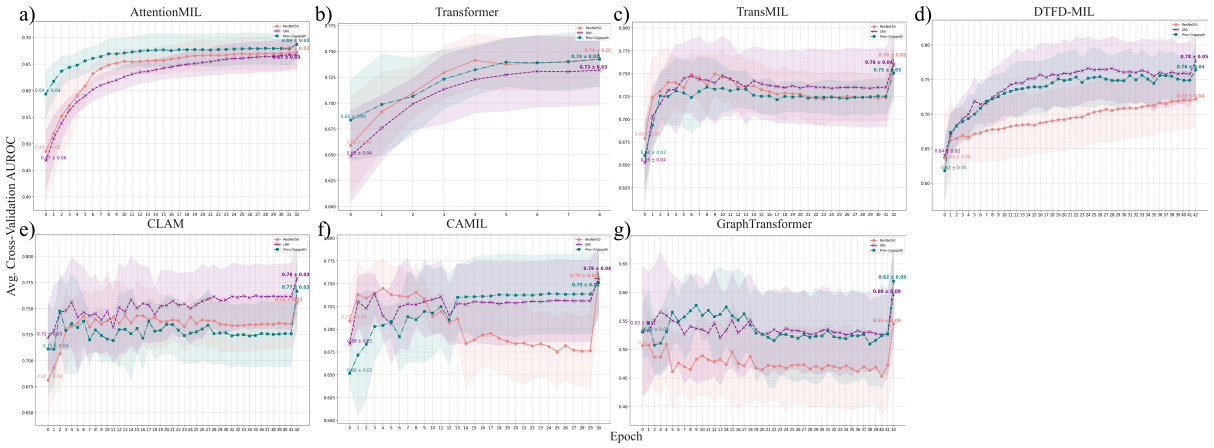


Figure 4.1: Average AUROC across folds per epoch shown per classifier. We label two "milestones", which is the average performance at the beginning of training, and in bold we show the highest mean cross-validation AUROC achieved at the end of training to illustrate the improvement brought by learning. 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 foundational model alternatives.

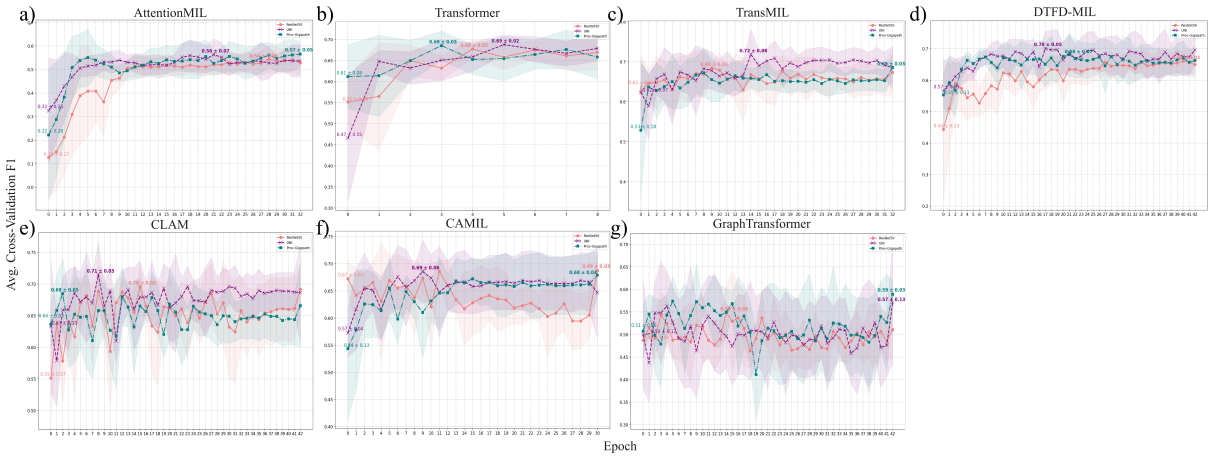


Figure 4.2: Average F1 across folds per epoch shown per classifier. We label two "milestones" in the same manner as in Figure 4.1, where we observe that the highest mean cross-validation F1 is not necessarily achieved at the end of training.

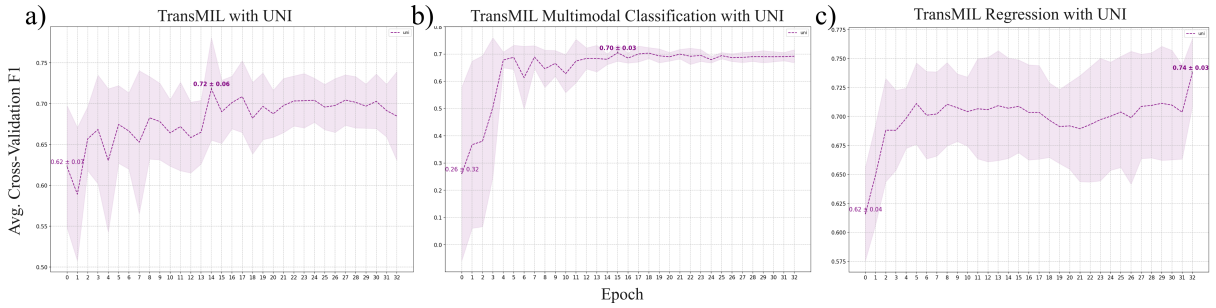


Figure 4.3: Average F1 across folds per epoch shown for ablations of TransMIL with UNI: TransMILMultimodal and TransMILRegression. We label two "milestones" in the same manner as in Figure 4.1. a) is the same lineplot as Figure 4.2c's UNI encoder. b) Interestingly, the average scores across folds is more stable, suggesting that multimodal fusion stabilizes training across folds.

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.

Our evaluation over the test set in terms of AUROC (Figure 4.4) and F1 (Figure 4.5) suggests that MIL algorithms trained with foundational feature encoders may lead to better generalization performance. This is explained by how the embeddings $\mathcal{W}$ computed from UNI and Prov-Gigapath are semantically richer than their ResNet-50 counterpart, which facilitate each MIL model learning an association between input features and output. Here, "semantically richer" encapsulates the idea that foundational models' features capture complex patterns . We further note that since we're employing open-source TCGA slides, our colorectal WSIs could have been sourced for training the foundational models.

We also compute the test F1 over ablations of TransMIL using the UNI feature encoder and report results in Figure 4.6, 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 colorectal 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 4.7. 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 4.12.

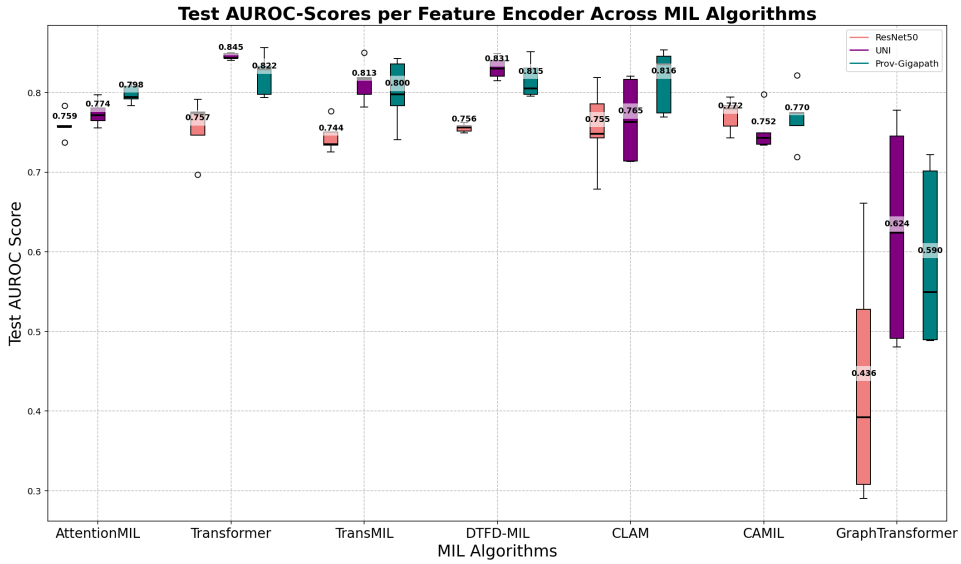


Figure 4.4: Average test AUROC obtained from the 5 independent optimal models per fold, per feature encoder. We notice that for all algorithms, except CAMIL, at least one of both feature encoders surpasses the performance of the ResNet50 encoder, albeit with overlapping std. errors (i.e. in the figure the purple and teal bars are often higher than their lightcoral counterpart). This suggests that the choice of foundational feature encoders helps with generalization.

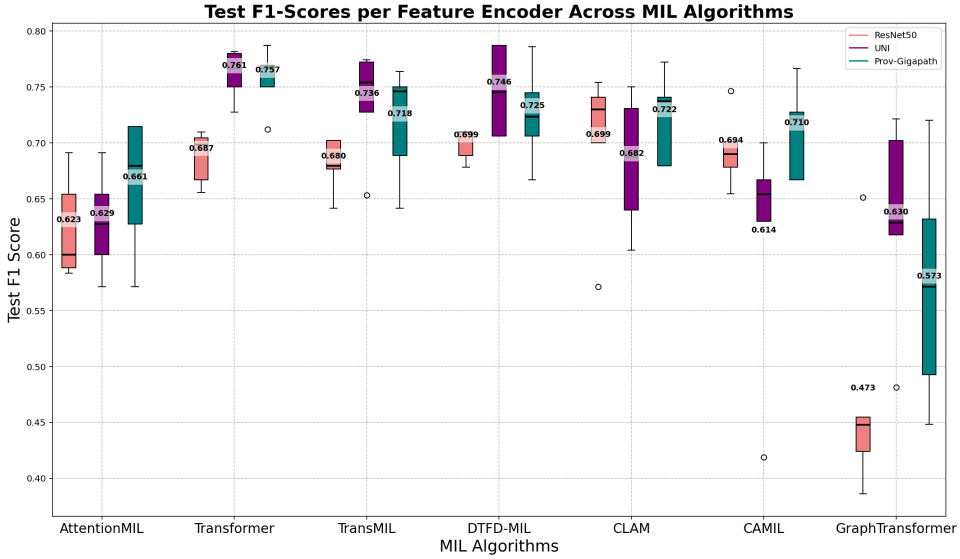


Figure 4.5: Boxplot of test F1 scores obtained from the 5 independent optimal models per fold, per feature encoder, evaluated over the test set. The often higher test scores (i.e. higher purple and teal bars) achieved by the UNI and Prov-Gigapath feature encoders suggests better generalization capabilities brought by foundational feature encoders in comparison to the standard ImageNet-pretrained ResNet50.

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 4.1. We also report an ensemble for each of TransMIL's ablations at Table 4.2. We generally observe higher scores than in Figures 4.4, 4.5, 4.6, 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

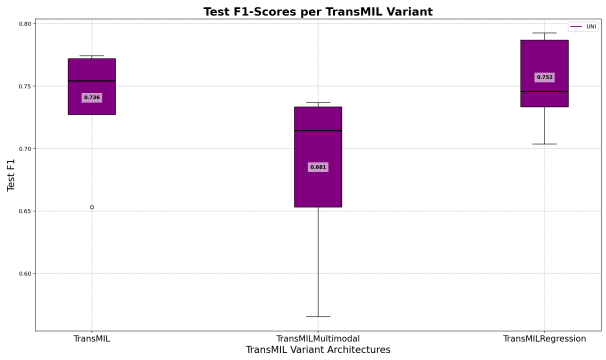


Figure 4.6: Boxplot of test F1 scores obtained from the 5 independent optimal TransMIL ablation models per fold, trained using the UNI encoder, evaluated over the test set.

ensembling helps correct a previously wrong prediction made by a single best TransMIL (Figure 4.8). 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 standard error intervals into their predictions.

For interested pathologists, we share a more comprehensive view of heatmaps generated by our Ensemble TransMIL in the Appendix Figure 6.1, spanning those generated in correct and wrong predictions, along with samples of patches where the ensemble bases its predictions on.

4.3 Multimodal fusion improves interpretability with some performance sacrifice

We performed a baseline experiment with only 27 clinical features with an ensemble of random forest binary classifiers (RFC) and obtain a test AUROC of 0.602 and test F1 of 0.528, which showcases that only looking at $\mathbf{f}$ is not enough to build an expressive classifier of G0-arrest. As such, relying on morphological features from the images instead help improve classification performance, as our best MIL algorithms can achieve higher

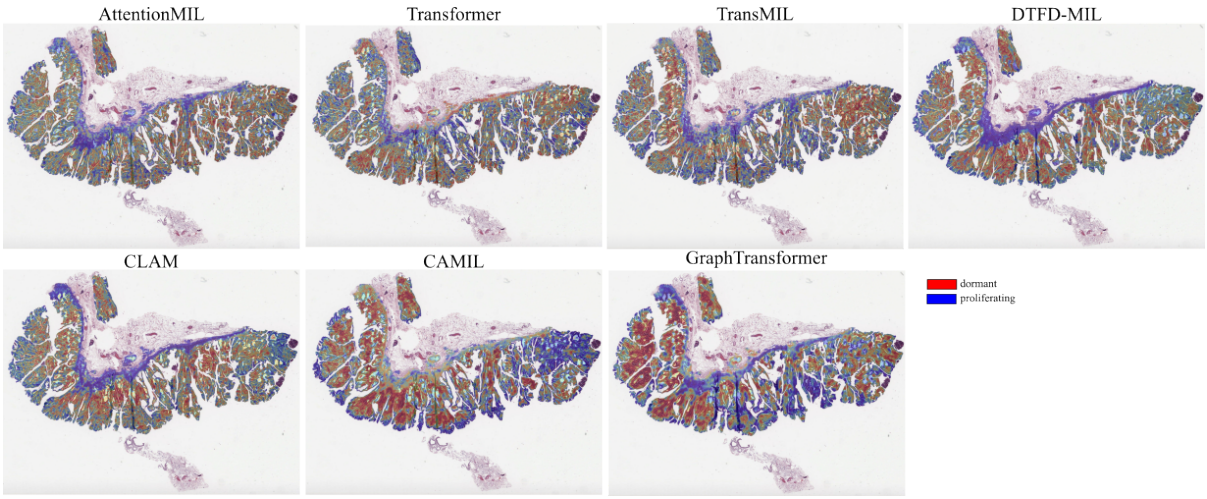


Figure 4.7: Side-by-side comparison of heatmaps generated by benchmarked algorithms using the UNI feature encoder. Each heatmap is obtained by mapping the average attention scores from an ensemble of the best models per CV fold. All heatmaps explain a TP prediction except for AttentionMIL and GraphTransformer 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.

AUROC and F1. The clinical features that the ensemble RFC learns to be the most important ones are show in Figure 4.9

This contrasts with the clinical features that the Ensemble TransMILMultimodal focuses on based on the IG values computed over the test set. We emphasize that both the feature importances and IG are fundamentally different methods, as such their values can only be interpreted within each method, but not compared with each other. For instance, the RFC's feature importances are computed during training, while IG is computed on a per-data-point basis and aggregated across the test set to observe on average how each feature influenced the prediction.

By analyzing the IG values, 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

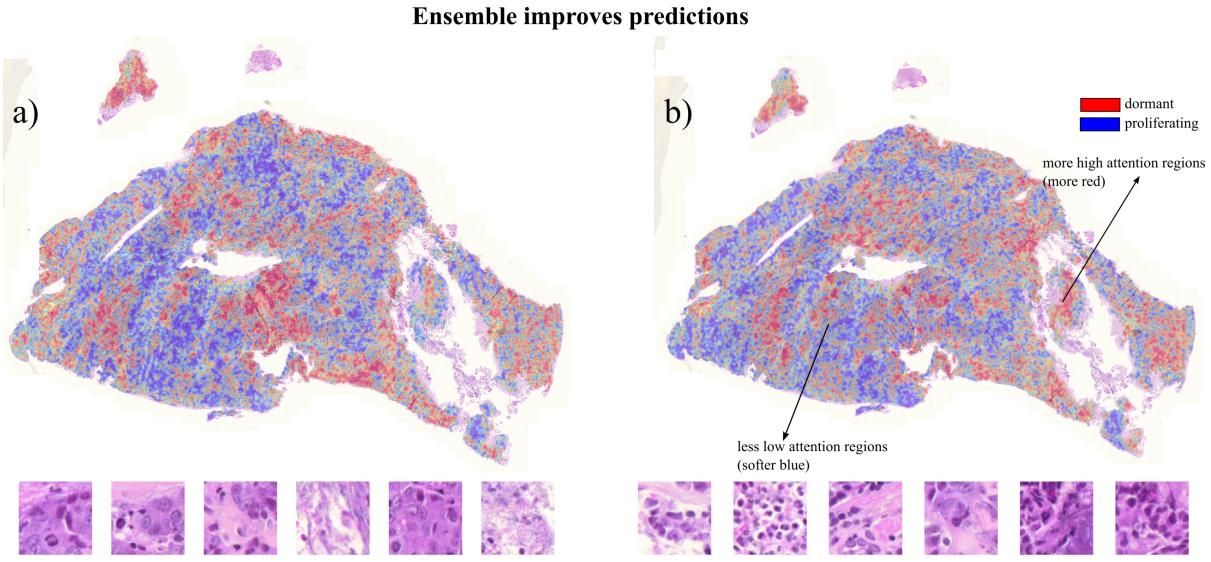


Figure 4.8: Depiction of how an ensemble improves predictions, and this is reflected in the heatmaps generated. At a) we show a single TransMIL trained with the UNI feature encoder making a false negative prediction on a testpoint. At b), this is corrected into a true positive prediction by an ensemble 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.

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, TransMILMulti-

	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 4.1: 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.

modal 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 TransMIL-Multimodal, coinciding with the ensemble RFC, identifies patient's age, gender and pre-operative 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.

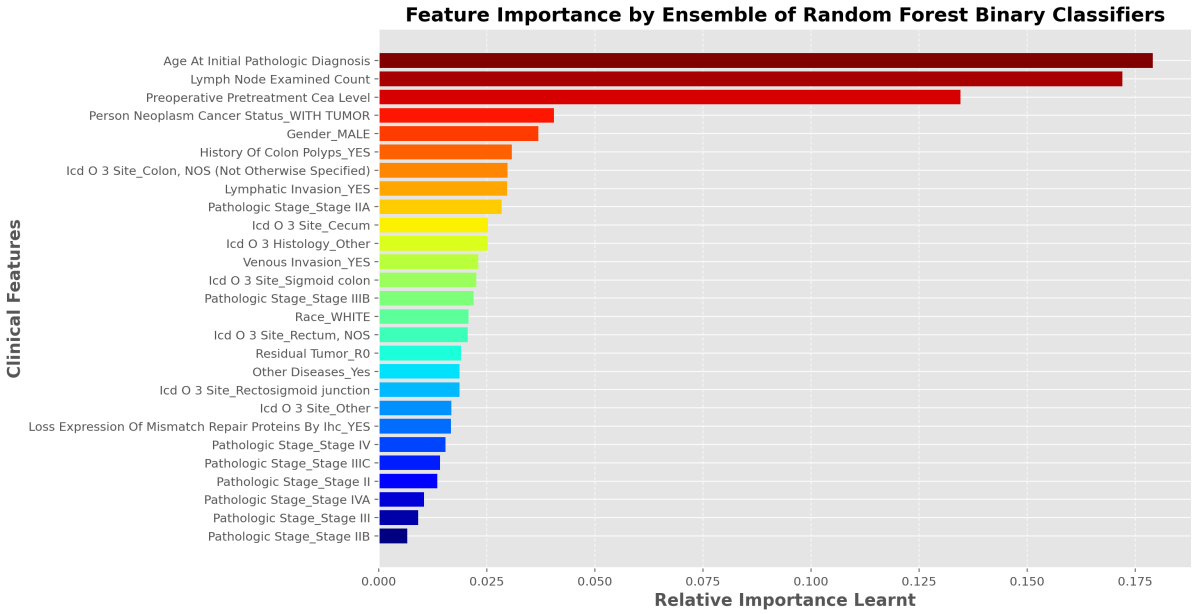


Figure 4.9: Descendently ranked relative importance values per clinical feature computed by averaging the feature importance scores obtained from the ensemble RFC's during training during 5-fold cross-validation. We employ the jet color map, whereby red indicates high relative importance scores, while blue otherwise. The theoretical range of the feature importance scores is in $[0, 1]$, and all the relative importance scores must sum to 1.

	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 4.2: 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 4.1. We obtain the highest F1 through binarizing regression scores.

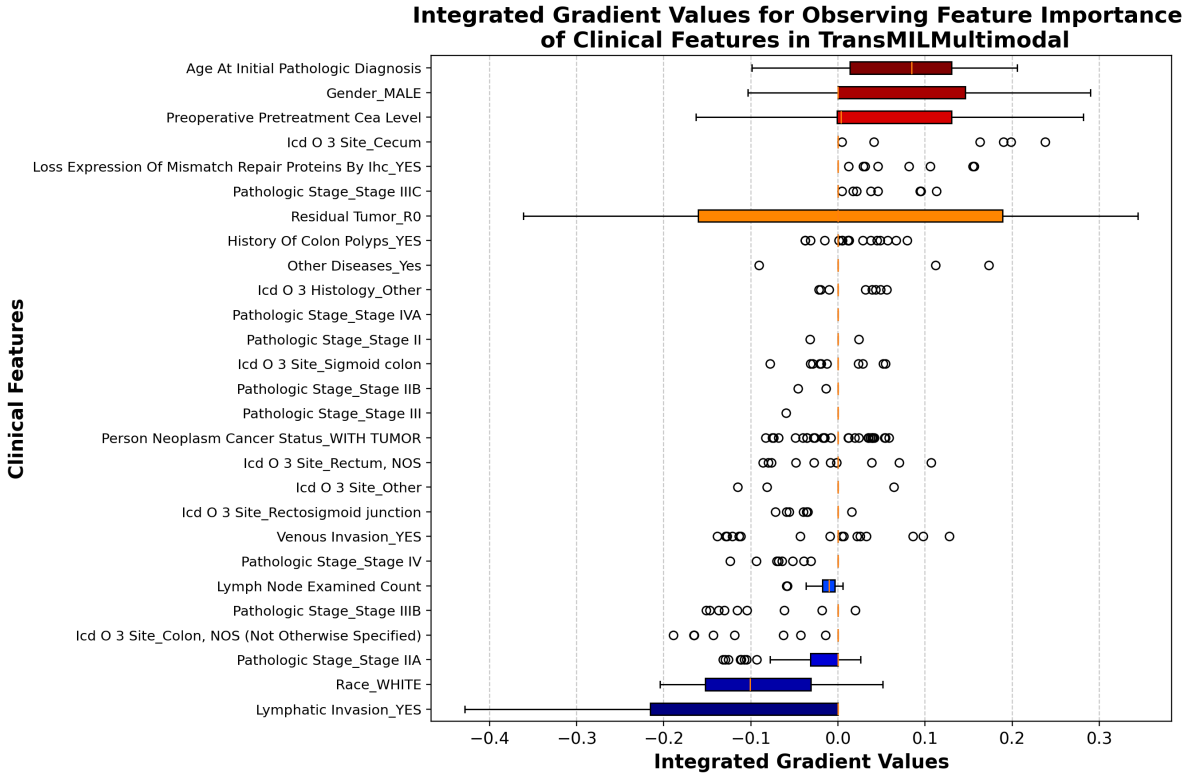


Figure 4.10: Descendently ranked IG values per clinical feature obtained by averaging the IG values obtained with an ensemble TransMILMultimodal predicting over the test set. We use the same color code as before, except that with IG values, the theoretical range extends to $[-\infty, +\infty]$, where 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.

4.4 Inductive biases yield more biologically meaningful predictions

Spatial context-awareness: from the heatmaps shown in Figure 4.7, 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

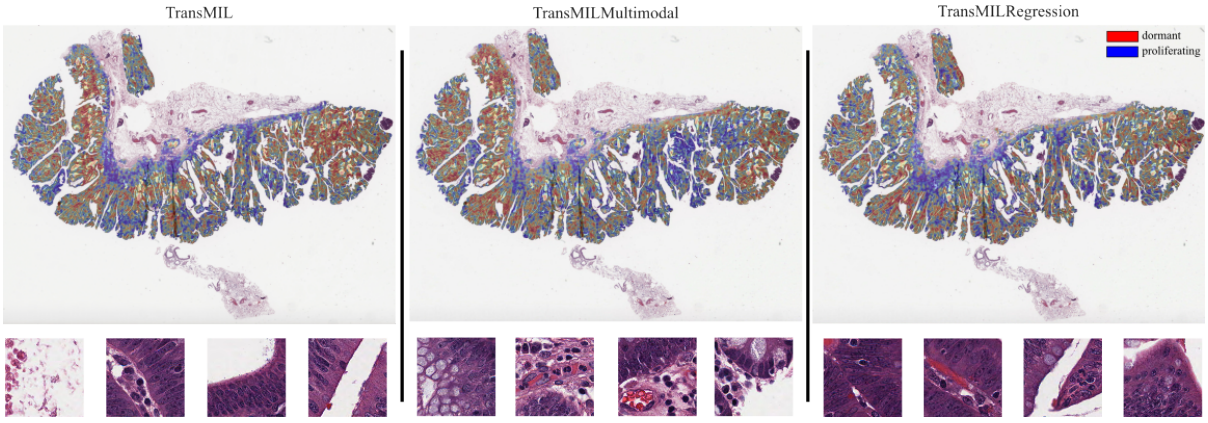


Figure 4.11: Side-by-side comparison of heatmaps generated by ablations of TransMIL with the UNI feature encoder. Below each heatmap is a sample of 4 patches with the highest attention scores contributing to the prediction of G0-arrest, and are all TP predictions. For TransMIL and TransMILMultimodal, 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

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 4.12). 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 4.2), 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 4.11, 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.

4.5 Limitations and future work

We have thoroughly benchmarked MIL algorithms, discussed the generalization capability brought by foundational feature encoders, the performance improvements brought by ensembling, analyzed the clinical features of the multimodal MIL model and how inductive biases help align MIL model outputs to biological expectations. We proceed in discussing the limitations of our work, and propose future research directions.

G0-arrest and tumor heterogeneity: Our main interest revolves around unveiling the tumor cells in G0-arrest in the colorectal TME. However, we note that our slide-level labels y are computed from bulk-RNA sequencing data, thus there is a mix of RNA-seq signals derived 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: One of the main nuances of our work revolves around evaluation. An interplay of small dataset, gigapixel sized WSI, challenging task of molecular-level prediction and constrained computational resources drive us to decisions with a few trade-offs. A small dataset results in a small test set (58 images) to guar-

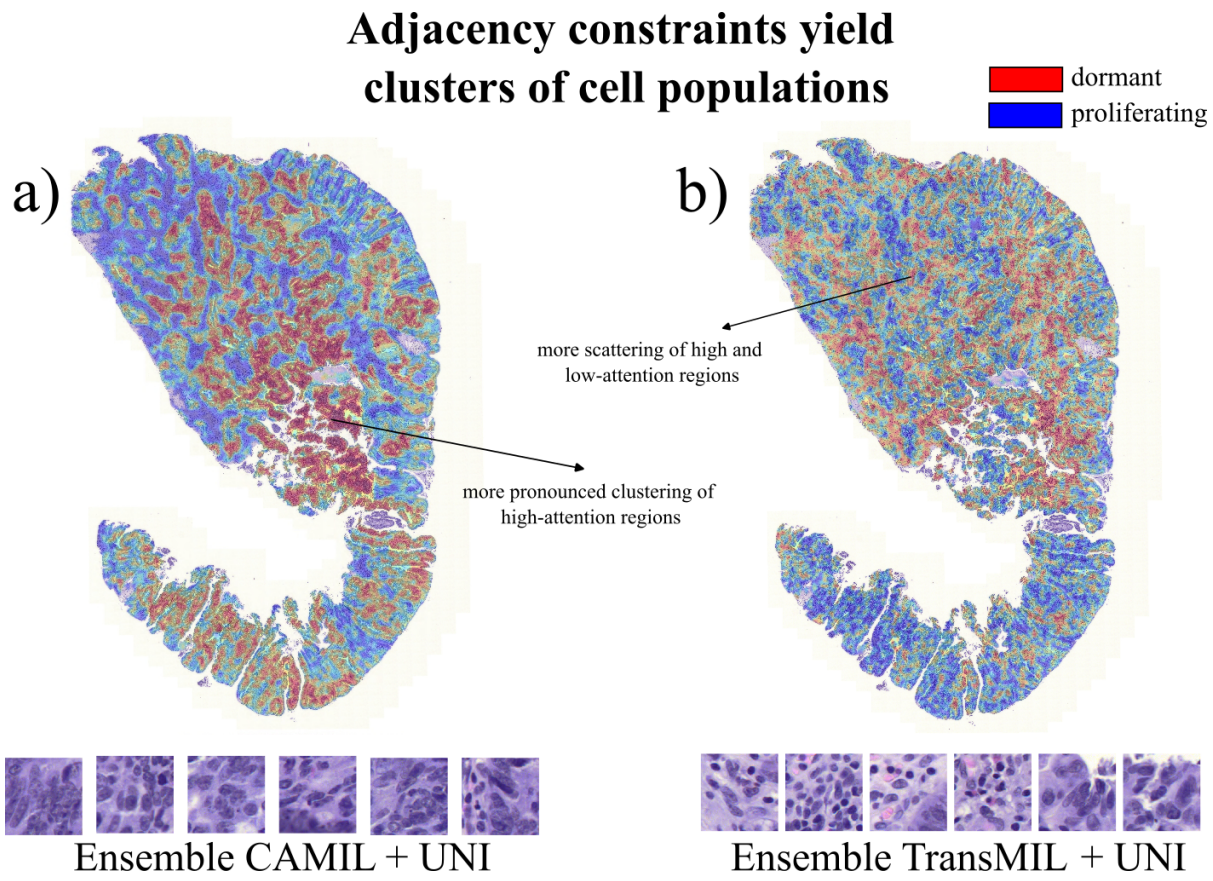


Figure 4.12: Adjacency constraints introduced via the graph representation of a WSI helps the model visualize more pronounced clusters of cell populations. While this comes at the expense of some performance loss, the spatially-constrained heatmaps produced by a) CAMIL and GraphTransformer align more 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.

antee model capturing important morphological features during training. Despite doing 5-fold cross-validation for each algorithm for a thorough evaluation of their generalization performance, there are more rigorous ways to evaluate each of our benchmarked models. Some research teams measure model generalization capability by making test-splits consist from a patient cohort such as TCGA stratified by different clinical sites [El Nahhas et al., 2024] or datasets belonging to different patient cohorts [Wagner et al., 2023],

e.g. they train their transformer model to predict MSI on colorectal WSI on cohorts such as TCGA, CPTAC, among others, except YCR-BCIP and test their model on biopsies from YCR-BCIP. Both methods correspond to out-of-distribution evaluation, which we aim in future work. Furthermore, regarding interpretability analysis, we aim to employ CRC tissue which has undergone ST or IHC analysis in order to obtain spatially-resolved ground truths, and evaluate the accuracy of the heatmaps produced by our models per feature encoder [Parreno-Centeno et al., 2022]. Such evaluation would help us answer questions such as whether the use of foundational 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 foundational 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 foundational 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 through self-supervised learning. This would greatly increase the size and heterogeneity of our dataset, which

allows us to perform more thorough evaluation. However, this 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 foundational 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 foundational 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*.

[Azizi et al., 2023] Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., Tomasev, N., Mitrovic, J., Strachan, P., Mahdavi, S. S., Wulczyn, E., Babenko, B., Walker, M., Loh, A., Chen, P.-H. C., Liu, Y., Bavishi, P., McKinney, S. M., Winkens, J., Roy, A. G., Beaver, Z., Ryan, F., Krogue, J., Etemadi, M., Telang, U., Liu, Y., Peng, L., Corrado, G. S., Webster, D. R., Fleet, D., Hinton, G., Houlsby, N., Karthikesalingam, A., Norouzi, M., and Natarajan, V. (2023). Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7:756–779.

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

[Bontempo et al., 2023] Bontempo, G., Porrello, A., Bolelli, F., Calderara, S., and Ficarra, E. (2023). Das-mil: Distilling across scales for mil classification of histological wsis. *Lecture notes in computer science*, pages 248–258.

[Campanella et al., 2019] Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25:1301–1309.

[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.

[Dapena et al., 2015] Dapena, C., Bravo, I., Cuadrado, A., and Figueroa, R. I. (2015). Nuclear and cell morphological changes during the cell cycle and growth of the toxic dinoflagellate alexandrium minutum. *Protist*, 166:146–160.

[de Haan et al., 2021] de Haan, K., Zhang, Y., Zuckerman, J. E., Liu, T., Sisk, A. E., Diaz, M. F. P., Jen, K.-Y., Nobori, A., Liou, S., Zhang, S., Riahi, R., Rivenson, Y., Wallace, W. D., and Ozcan, A. (2021). Deep learning-based transformation of h&e stained tissues into special stains. *Nature Communications*, 12.

[Eastwood et al., 2023] Eastwood, M., Sailem, H., Tudor Marc, S., Gao, X., Offman, J., Karteris, E., Montero Fernandez, A., Jonigk, D., Cookson, W., Moffatt, M., Popat, S., Minhas, F., and Lukas Robertus, J. (2023). Mesograph: Automatic profiling of mesothelioma subtypes from histological images. *Cell reports medicine*, 4:101226–101226.

[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-Bruchle, 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.

[Gao et al., 2017] Gao, X., Zhang, M., Tang, Y., and Liang, X.-h. (2017). Cancer cell dormancy: mechanisms and implications of cancer recurrence and metastasis. *OncoTargets and Therapy*, Volume 10:5219–5228.

[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.

[Hardy,] Hardy, M. The cell cycle. *slcc.pressbooks.pub*.

[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., GrÃ¼nwald, B., Egger, J., and Kleesiek, J. (2024). Cellvit: Vision transformers for precise cell segmentation and classification. *Medical image analysis*, 94:103143–103143.

[Huang et al., 2022] Huang, W., Hickson, L. J., Eirin, A., Kirkland, J. L., and Lerman, L. O. (2022). Cellular senescence: the good, the bad and the unknown. *Nature Reviews*

Nephrology, 18.

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

[Joanito et al., 2022] Joanito, I., Wirapati, P., Zhao, N., Nawaz, Z., Yeo, G., Lee, F., Eng, C. L. P., Macalinao, D. C., Kahraman, M., Srinivasan, H., Lakshmanan, V., Verbandt, S., Tsantoulis, P., Gunn, N., Venkatesh, P. N., Poh, Z. W., Nahar, R., Oh, H. L. J., Loo, J. M., Chia, S., Cheow, L. F., Cheruba, E., Wong, M. T., Kua, L., Chua, C., Nguyen, A., Golovan, J., Gan, A., Lim, W.-J., Guo, Y. A., Yap, C. K., Tay, B., Hong, Y., Chong, D. Q., Chok, A.-Y., Park, W.-Y., Han, S., Chang, M. H., Seow-En, I., Fu, C., Mathew, R., Toh, E.-L., Hong, L. Z., Skanderup, A. J., DasGupta, R., Ong, C.-A. J., Lim, K. H., Tan, E. K. W., Koo, S.-L., Leow, W. Q., Tejpar, S., Prabhakar, S., and Tan, I. B. (2022). Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nature Genetics*, 54:963–975.

[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.

[Mitra et al., 2018] Mitra, M., Ho, L. D., and Coller, H. A. (2018). An in vitro model of cellular quiescence in primary human dermal fibroblasts. *Methods in molecular biology (Clifton, N.J.)*, 1686:27–47.

[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.

[Oki et al., 2014] Oki, T., Nishimura, K., Kitaura, J., Togami, K., Maehara, A., Izawa, K., Sakaue-Sawano, A., Niida, A., Miyano, S., Aburatani, H., Kiyonari, H., Miyawaki, A., and Kitamura, T. (2014). A novel cell-cycle-indicator, mvenus-p27k-, identifies quiescent cells and visualizes g0-g1 transition. *Scientific Reports*, 4.

[Organization, 2023] Organization, W. H. (2023). Colorectal cancer.

[Pack et al., 2019] Pack, L. R., Daigh, L. H., and Meyer, T. (2019). Putting the brakes on the cell cycle: mechanisms of cellular growth arrest. *Current Opinion in Cell Biology*, 60:106–113.

[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., MÃ¼ller, 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.

[Santos-de Frutos and Djouder, 2021] Santos-de Frutos, K. and Djouder, N. (2021). When dormancy fuels tumour relapse. *Communications Biology*, 4:1–12.

[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.

[Siegel et al., 2023] Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A., and Jemal, A. (2023). Colorectal cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, 73.

[Smith et al., 2024] Smith, K. D., Prince, D. K., MacDonald, J. W., Bammler, T. K., and Akilesh, S. (2024). Challenges and opportunities for the clinical translation of spatial transcriptomics technologies. *PubMed*, 4:49–63.

[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*.

[Stahlschmidt et al., 2022] Stahlschmidt, S. R., Ulfenborg, B., and Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23.

[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.

[UK, 2015] UK, C. R. (2015). Bowel cancer incidence statistics.

[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.

[Vorontsov et al., 2023] Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Mathieu, P., van Eck, A., Lee, D., Viret, J., Robert, E., Wang, Y. K., Kun, J., Le, M., Bernhard, J., Godrich, R., Oakley, G. J., Millar, E. K., Hanna, M. G., RetÃ¡mero, J. A., Moye, W. A., Yousfi, R., Kanan, C., Klimstra, D. S., Rothrock, B., and Fuchs, T. J. (2023). Virchow: A million-slide digital pathology foundation model. *arXiv (Cornell University)*.

[Wagner et al., 2023] Wagner, S. J., ReisenbÃ¼chler, 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., GÃ¶genÃ¼r, 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.

[Wang et al., 2022] Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., and Han, X. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559.

[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., 2024a] 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. (2024a). 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.

[Xiao et al., 2024b] Xiao, J., Yu, X., Meng, F., Zhang, Y., Zhou, W., Ren, Y., Li, J., Sun, Y., Sun, H., Chen, G., He, K., and Lu, L. (2024b). Integrating spatial and single-cell transcriptomics reveals tumor heterogeneity and intercellular networks in colorectal cancer. *Cell Death & Disease*, 15:1–11.

[Xu et al., 2024] Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., GonzÃ¡lez, 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.

[Yacob et al., 2023] Yacob, F., Siarov, J., Villiamsson, K., Suvilehto, J. T., SjÃ¶blom, L., Kjellberg, M., and NeittaanmÃ¤ki, N. (2023). Weakly supervised detection and classification of basal cell carcinoma using graph-transformer on whole slide images. *Scientific Reports*, 13.

[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*.

6 | Appendix

6.1 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.
- 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

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)

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 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 $n_{iter} = 100$ times, and we get rid of 3 features 'Pathologic Stage IIC', 'Pathologic Stage IIIA', and 'Pathologic Stage IVB' which for more than threshold = 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.

6.2 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 6.1: 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.

6.3 Interpretability analysis of Ensemble TransMIL with UNI feature encoder

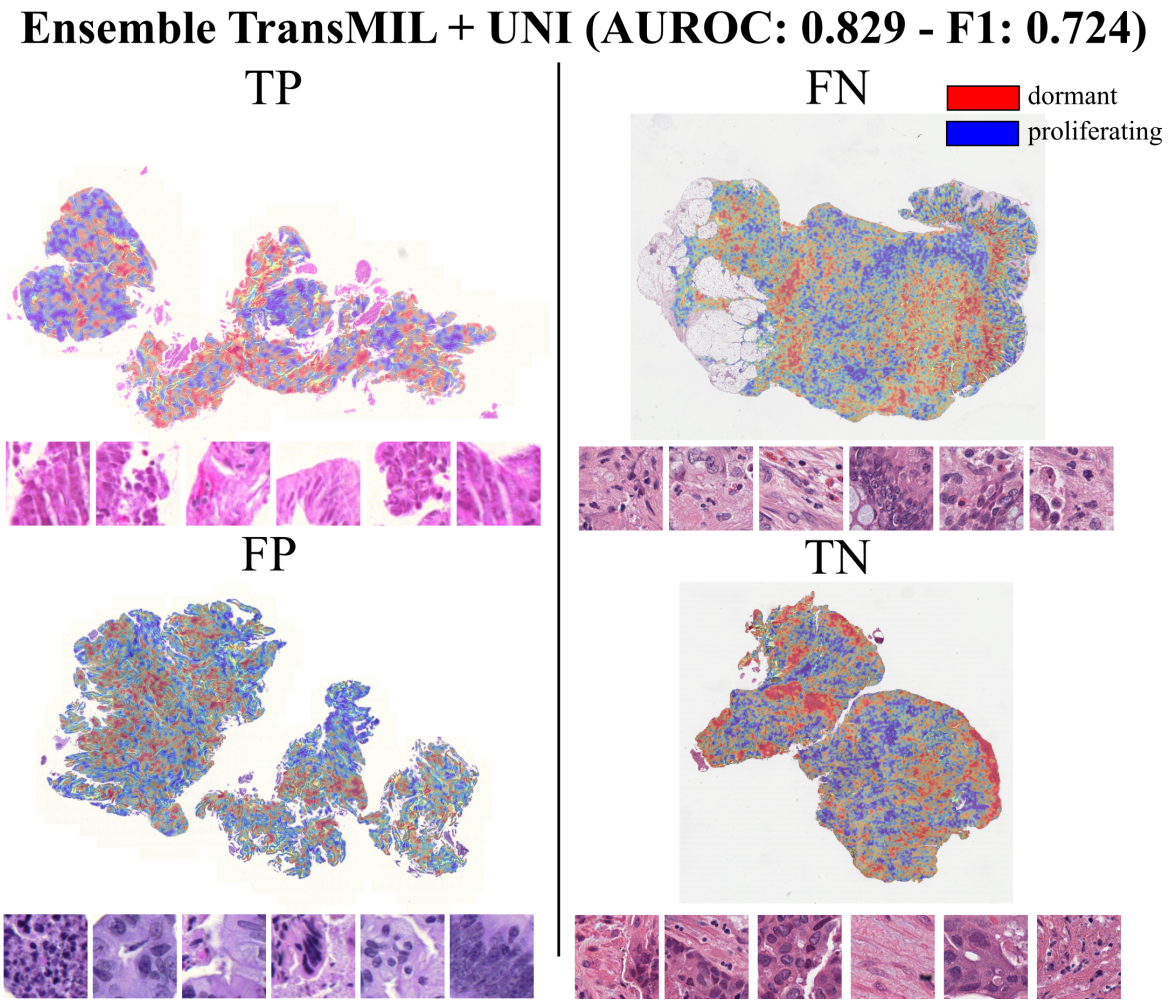


Figure 6.1: Heatmaps generated by the Ensemble TransMIL with the UNI feature encoder. We provide correct and incorrect classifications, and below each heatmap we append a sample of 6 patches according to their attention scores contributing to the slide-level prediction. For TP and FP, these patches have the highest attention scores explaining a positive prediction. For TN and FN, the patches have the lowest attention scores explaining a negative prediction.