MULTI-MODAL DATA ANALYSIS FOR PATIENT OUTCOME PREDICTION IN COLORECTAL CANCER

by

Kexin Ding

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2024

Approved by:

_____
Dr. Aidong Lu

_____
Dr. Shaoting Zhang

_____
Dr. Min Shin

_____
Dr. Yaorong Ge

ABSTRACT

KEXIN DING. Multi-modal Data Analysis for Patient Outcome Prediction in Colorectal Cancer. (Under the direction of DR. AIDONG LU)

Understanding and characterizing cancer patient outcomes is challenging and involves multiple clinical measurements (e.g., imaging and genomics biomarkers). Enabling multimodal analytics promises to reveal novel predictive patterns that are not available from singular data input. In particular, exploring histopathological and genomics sequencing data provides a synergistic path to understanding the deep insights of cancer biology. In this dissertation, we first present a graph-based neural network framework that allows multi-region spatial connection of tiles to predict molecular profile status in colorectal cancer. We demonstrate the validity of spatial connections of tumor tiles built upon the geometric coordinates derived from the raw histopathological images. These findings capture the interaction between histopathological characteristics and a panel of molecular profiles of treatment relevance. Second, we propose a multimodal transformer integrating pathology and genomics insights into colorectal cancer survival prediction. The proposed unsupervised pretraining captures the intrinsic interaction between tissue microenvironments in WSI and a wide range of genomics data (e.g., miRNA-sequence, copy number variant, and methylation). After the knowledge aggregation in pretraining, the task-specific model finetuning expands the scope of data utility applicable to both multi- and single-modal data. Finally, we introduce a contrastive pathology-and-genomics pretraining to enhance patient survival prediction by extracting the multimodal interaction for each patient while distinguishing the differences among various patients. This dissertation provides solutions for addressing the challenges in understanding multimodal disease data, leading to improved overall performance of patient outcome prediction in colorectal cancer.

ACKNOWLEDGEMENTS

I have almost finished my Ph.D. study journey with the support of many people in my life. First and foremost, I would like to thank my advisor, Dr. Aidong Lu. You helped and supported me a lot in the past five years. You tried your best to assist me in dealing with each difficulty that I met in each of the milestones to receive my PhD degree. I also want to thank my co-advisor, Dr. Shaoting Zhang. You provided me with the best research environment and support so that I could focus on my research without any concerns. Thank you also for reading multiple drafts of papers and helping me understand the histopathological image analysis. Further, I want to show my gratitude to Dr. Min Shin and Dr. Yaorong Ge for being on my committee and providing their support to my dissertation.

I sincerely want to thank my mentor, Dr. Mu Zhou. We had a close and perfect collaboration in the past five years. I am grateful that he tried his best to lead me to walk into the world of research. He taught me how to be a Ph.D. student, set up and finish a project, write a paper, and even find my first job. He showed me his biggest responsibility and patience. He watched each of my steps in my Ph.D. study. He would say congrats on my little progress, and he would never give up on me even though I was in a bad research situation. For each of my projects, we discussed the idea and read the paper together. When I suffer technical difficulties, he always leads me to find the solution by myself instead of telling me the answer directly. I learned a lot from him, including the approach and the attitude of being a researcher. I will keep the research habit as his, no matter whether I go to the industry or academia in the future. In this acknowledgment, I want to show my highest respect and thank him for his support and help in the past days, months, and years.

I want to thank my best friend, Yue Lyu, who is currently working at Expedia. She encouraged me in the past five years and walked together with me when I was in the dark. I can always remember her kindness and support during my entire PhD study.

Several times, I felt upset and hopeless, but she encouraged me and tried to help me. She is my best and most supportive friend, and I hope that our friendship can continue forever. I also want to thank my lab mates. I will never forget their kindness when I came to the lab for the first time. They invited me to join the regular lab meetings and research club activities. Although we have not met each other frequently since the pandemic, I really enjoyed my time in the lab. Many thanks to Dr. Min Shin for organizing the lab so well. I also want to thank my friend Hehuan Ma, who brought so much warmth and happiness while I worked remotely in Texas during the pandemic.

Furthermore, I want to thank my family, parents, and husband. In each dark time, I could always feel their support and love. My parents never pressured me and always supported me in doing anything I wanted. In the past twenty-eight years, they have provided superior living and education conditions so I could have a safe and happy life. They always support me in pursuing what I want and feeling satisfied even if I am not the best one. My husband and I have lived together for a long time since the pandemic, and he is also a Ph.D. student of the same age as me. However, he took on most of the housework and responsibility of our family, so I only needed to focus on my research. When we watched the movie "Anatomy of A Fall", I felt so shocked that the arguments in this movie are so familiar to me. In my world, he is such a clever, sincere, kind, and reliable friend to me. We have faced and overcome so many drawbacks and become the strongest backup to each other. We have been together for nine years so far and have been to several wonderful places in the world. These nine years are the perfect days for me. Thanks for the past few years that we grew up together, and I am looking forward to walking together with him in later years.

Finally, I want to acknowledge Springer Nature and Elsevier Journals, which grant me the right to reuse the version of the record or any part of my previous publications in my dissertation. The text content, figures, and tables in Chapter 3 are reproduced from my previous publication [1] with permission from Elsevier. The

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Background and Motivation

Colorectal cancer ranks as the third most common cancer globally, with approximately 10% of all newly diagnosed cancers and around 9.4% of all cancer-related deaths in 2020 [3]. The five-year survival rate of colorectal cancer is 67%, meaning that patients with colorectal cancer can likely survive more than five years with proper treatment. Timely diagnosis and proper treatment decision-making become crucial factors in improving patient survival rates. Targeted therapy is a type of cancer treatment that is highly related to molecular alteration and utilizes drugs to attack specific molecular targets, such as proteins in cancer cells. These treatments have demonstrated their capability for patient treatment without affecting normal cells. Such treatments for improving patient survival ratio require a deep understanding of genetic changes and proteins that drive cancer, which is beneficial to the selection of targeted therapy that works best against a particular type of cancer.

Benefits from the rapid evolution of gene sequencing technology, genomics profiles (e.g., miRNA-sequence) are widely used for regulating patient cancer progression and treatment [4, 5, 6, 7, 8, 9]. For instance, genome-wide molecular portraits play an important role in patient prognostic assessment and treatment decision-making [10]. Multi-level molecular characteristics are able to show spatial differences within the tumor tissue microenvironment [11]. Specific molecular profile alterations, such as KRAS mutation, are known as the driver gene related to cancer progression and are strongly associated with patient therapy in colorectal cancer [12]. Microsatellite instability (MSI), characterized by defective DNA mismatch repair (MMR) systems, provides key insights for colorectal cancer prognostic [13]. Additionally, protein an-

alytics can broaden the landscape of cancer genomics for various biomarker discovery [14]. In summary, these multi-scale molecular-wise biomarkers can deepen our understanding of cancer evolution, enabling better patient stratification and treatment [12]. Spatially intertwined regions within tumor tissues, harboring molecularly distinct features, indicate the existence of intra-tumoral heterogeneity [11]. Such tumor spatial heterogeneity shows the diverse distribution of tumoral molecular subpopulations, reflecting varying levels of sensitivity to treatment decision-making [15]. Evidence of spatial heterogeneity is primarily derived from transcriptional and genetic profiles obtained through physically isolated biopsies from a single tumor [1].

The growth of digitalized histopathological images becomes a valuable resource that enables rapid and precise cancer diagnosis, patient outcome understanding, and treatment decision-making. These images can capture extensive and detailed pathological patterns of disease that are not available in other modalities of medical images, such as ultrasound images and magnetic resonance imaging (MRI). The high-resolution images can provide a unique avenue to assess the spatial context of the entire tumoral microenvironment (e.g., cancer cells and their surrounding tissues) and tissue interactions. The rich tissue characteristics are highly related to patient disease assessment [16, 17]. Furthermore, the disease can be triggered by histopathological changes associated with key molecular variations, such as genetic mutations, copy number alterations, and protein expressions [12]. With heterogeneous disease causes, understanding the complex interactions between histopathological and genomic biomarkers in tumor tissue environments becomes a promising direction for enhancing patient outcomes in colorectal cancer [18]. For instance, cancer-related genotypes can manifest as histopathological phenotypes in images, which can be evaluated by pathologists for precise patient outcome predictions [18]. Histopathological images have demonstrated their unique benefits for enhancing prognostic biomarker prediction by exploring tissue microenvironment features. Such prognostic biomarkers are highly

related to target therapy treatment decision-making for improving patient survival. The identification of unique histopathological patterns that are sensitive to the underlying molecular mechanisms is crucial to improving our biological understanding and making more informed diagnoses. The conventional deep-learning methods, i.e., convolutional neural networks, have demonstrated their superior capabilities for image-based feature discovery [19, 20, 21]; however, these methods are unable to directly characterize the underlying spatial information of tumoral sub-regions and their interactions. Graph convolutional neural networks (GCNs) open the possibility of quantitative WSI integrating regional and spatial contexts in depth in terms of associations with cancer molecular signatures[22, 1, 23]. Considering the tumor microenvironment with strong regional differences in image contents, the interactions of image tiles are key to understanding the status of molecular outcomes. GCNs provide a viable path to discover differential spatial characteristics from histopathology to help assess molecular variation, patient outcome, and targeted therapy for patients with colorectal cancer.

While single-modal data (either imaging or genomics) has demonstrated its clinical significance, there have been limited efforts to leverage the joint multimodal information between cancer morphology (e.g., histopathological image) and molecular biomarkers (e.g., genomics sequencing data). In a broader context of patient assessment, evaluating cancer prognosis inherently involves a multimodal task that expects to integrate pathological and genomic insights. Hence, integrating multimodal knowledge can facilitate a deeper understanding across different scales, leading to the improvement of patient prognosis. The primary objective of fusing data in multiple modalities is to exploit modality-complementary knowledge among various modalities [24]. Supervised methods [25, 26, 27] have explored the feasibility of fusing multimodal data, including both image and non-image biomarkers. Conventional fusion strategies such as the Kronecker product can explore the complex

interactions between WSI and genomic characteristics for predicting patient survival outcomes [25, 26]. Alternatively, approaches like the co-attention transformer [27] can investigate genotype-phenotype interactions for patient prognostic understanding. However, these supervised methods are constrained by feature generalization issues and a heavy reliance on data annotation, which potentially increases the burden of human efforts. To mitigate the need for labeled data, unsupervised data fusion assesses the intrinsic relationships among multimodal representations for effective knowledge integration. For instance, the knowledge fusion among histopathological images, genomics sequencing data, and patient tabular clinical data can be achieved by unsupervised modality relevance calculations [24]. To expand the applicability of data, a study [28] developed a two-stage workflow to explore the multimodal information to guide the single-modal model for glioma grading. The workflow utilizes pathological and genomic information by training a multimodal teacher model firstly. Then, the pathology-only student model can distill multimodal knowledge from the teacher model and be optimized by the specific single-modal knowledge in the second stage. There is a growing recognition that the flexibility of data modality in model finetuning can broaden the application scenario of multimodal learning. Moreover, unlike natural vision-language datasets, the volume of multimodal medical datasets is not extensive enough, which increases the need to develop data-efficient analytics.

1.2 Contribution

In this dissertation, we first introduced a graph neural network approach that emphasizes the spatialization of tumor tiles toward a comprehensive evaluation of predicting cross-level molecular profile alterations from whole-slide images. Second, we developed a multimodal framework (i.e., **PathOmics**) for survival outcome understanding on pathological and genomics data. Finally, we designed a contrastive learning-based pathology-and-genomics multimodal framework for enhancing survival prediction. To summarize, our work of deep learning-based high-throughput analysis

for histopathological image analysis has the following contributions, which will be elaborated on in each chapter:

- We developed a graph neural network approach, highlighting the spatial representation of tumor tiles, aiming for a comprehensive assessment of predicting multi-scale molecular profile alterations, including a wide range of gene mutations, copy number alterations, and the level of protein expression from whole-slide images. To address the spatial heterogeneity inherent in colorectal cancer, we introduced a transformation strategy that converts whole-slide images (i.e., grid-structured data) to graph-structured data. We developed and evaluated the performance of our model on The Cancer Genome Atlas colorectal adenocarcinoma (TCGA-COAD) dataset, and its validation was conducted on two external datasets: The Cancer Genome Atlas rectum adenocarcinoma (TCGA-READ) and Clinical Proteomic Tumor Analysis Consortium colorectal adenocarcinoma (CPTAC-COAD). Additionally, we conducted predictions for microsatellite instability and provided result interpretability.

- We introduced a multimodal framework (i.e., **PathOmics**) to explore the interaction among pathology-and-genomics patterns for survival outcome assessment. Our contributions are summarized as follows. (1) Unsupervised Multimodal Pretraining: We leverage unsupervised pretraining to capture interactions between morphological and molecular biomarkers. We bridge the gap of modality heterogeneity by projecting multimodal embeddings into a shared latent space through relevance evaluation. The pretrained data fusion facilitates unique cross-modal pattern extraction using relevance-guided modality fusion. (2) Flexible Modality Finetuning: Our framework is able to combine the benefits of unsupervised pretraining and supervised finetuning data fusion. Task-specific finetuning could broaden the dataset utility by easily adapting the model with both single- and multi-modal data scenarios. (3) Data Efficiency with Limited

Data Size: Even with fewer finetuned data, our approach achieves a comparable performance, demonstrating efficiency compared to using the entire finetuning dataset(e.g., only 50% of the finetuned data).

- We designed a contrastive learning-based pathology-and-genomics framework (i.e., C-PathOmics) for enhancing multimodal survival outcome understanding. Our main contribution focuses on the unsupervised contrastively multimodal pertaining. Our contrastively unsupervised pretraining aims to fuse the multimodal data, enabling the exploration of inherent relevance between morphological and genomics biomarkers. To address the disparity in modality between histopathological images and genomic sequencing data, we employ a method that involves mapping the embeddings from each modality into a shared latent space by exploring the relevance between the embeddings from different modalities. In the latent space, we are able to achieve multimodal embedding, which could provide complementary modality information for enhancing patient outcome assessment. To evaluate the modality relevance, we developed a multimodal contrastive relevance evaluation to learn the relevance between different modalities in a single patient and distinguish the differences among different patients. The pre-trained model provides a unique path to utilizing relevance-guided modality data fusion, allowing the extraction of cross-modal patterns characterized by unique modality features from patients.

1.3 Dissertation Outline

We organized this dissertation as follows: Chapter 2 reviews relevant studies using histopathological image genomics data for patient outcome analysis. Chapter 3 presents an approach for utilizing the global context and spatial information of the histopathological image by utilizing a graph convolutional network for a large-scale molecular profile prediction on colorectal cancer. Chapter 4 introduces a multimodal

analysis for patient survival prediction via extracting the interaction of histopathological image and genomics non-image data. Chapter 5 proposes a contrastive learning-based multimodal method for enhancing survival prediction. Chapter 6 concludes the dissertation and discusses the future direction of the field.

CHAPTER 2: RELATED WORKS

2.1 Histopathological Imaging Enables Molecular Profile Predictions

In histopathological imaging, the diagnostic-related phenotypic alterations in tumor cells and their microenvironment can inherently be caused by molecular changes [29]. Properly utilizing image-based biomarkers for molecular profile alteration understanding can help us explore the interaction that reveals the characteristics of cancer among diverse data modalities. Several efforts have been made to explore associations between histopathological images and genomic data in cancer research. With advancements in genomics data accessibility, the cost and time constraints associated with genomic analysis have considerably reduced in recent years. Consequently, deep learning models have performed an important role in revealing complex relationships among tissue morphological characteristics, biomarkers, and disease diagnosis.

Convolutional neural networks (CNNs) successfully achieved good performance by using histopathological images to predict molecular profiles in colorectal cancer. A widely used workflow involves convolutional feature extraction and tile weight determination for outcome prediction [19]. Schmauch et al. [30] extended an image-based CNN model to predict RNA-based transcriptomic profiles. With model finetuning, they reported AUC 81% on MSI classification on the TCGA-CRC-DX. Recently, a deep learning framework[20] explored the association between cellular composition profile and molecular profiles for the molecular pathway and key mutation predictions in colorectal cancer. They use both four-fold cross-validation (AUC 86%) and train-test splits (AUC 90%) for the MSI prediction performance on the TCGA-CRC-DX cohort. Additionally, with train-test splits of the dataset, they also reported the TP53 mutated and wildtype prediction (AUC 73%) and KRAS mutated and

wildtype prediction (AUC 60%). A clinical-grade CNN model is developed to predict microsatellite instability on histopathological slides on colorectal cancer [31]. By using three-fold cross-validation, they achieved the performance of MSI prediction on the TCGA-CRC-DX cohort (e.g., the combination of TCGA-COAD and TCGA-READ datasets) with an AUC value of 74%. Under the scenario of limited data, the deep learning model still showed superior performance in molecular profile prediction. For instance, MSINet [32] was proposed for classifying MSI status using a small number of whole-slide images (WSIs). Additionally, to better leverage limited patient annotations for robust model training, weakly supervised methods have been widely used in predicting molecular profile alterations using histopathological slides. With achieving good performance on various types of molecular profile prediction tasks, the weakly supervised architecture [20] has demonstrated its capability for assisting automatically cancer understanding. The ResNet18 model was used to classify tumor and non-tumor regions, enabling tumor region patch selection, which was then utilized for tuning an adapted ResNet34 for molecular alteration prediction. The possibility score of specific gene mutation classes of tumor region tiles predicted by the model. Then, a pretrained cell nuclei segmentation and classification tool is applied to the identified tiles, which are ranked and selected by their possibility score. Beyongd focusing a specific cancer, pan-cancer studies are emerging because of the complex interactions among multiple cancers. These studies [33, 29] utilized CNN-based frameworks with transfer learning to extract image tile features and achieve patient-wise genetic profile prediction by the average prediction results among tiles. Kather [29] evaluated multiple tumor types in the TCGA dataset to predict molecular and genomic subtypes, point mutation, and hormone receptor status from histopathological mage images. Fu [33] extended a relative study across multiple types of cancers in the TCGA dataset. In the TCGA-COAD cohort, they reported five-fold cross-validated AUC values ranging from 59.61% to 72.02% for APC, TP53,

KRAS, PIK3CA, SMAD4, and FBXW7mutations. In the TCGA-READ cohort, they reported cross-validated AUC values ranging from 50.21% to 74.79% for KRAS, APC, TP53, PIK3CA, FBXW7, and SMAD4 mutations.

Despite the promising performance of CNN models in predicting biomarkers enabling better cancer understanding, there remains limited exploration of the topological structure within tissue microenvironments. Graph neural networks (GNNs) offer a novel approach to linking histopathological images with molecular outcomes, although they have not been extensively explored in this context. Notably, a study utilized cell-based graph analysis to predict the HER2 and PR status of breast cancer, which relies on a dependency on extra cell detection and neighborhood clustering [34]. HoverNet has emerged as a popular model for nuclei segmentation and classification, facilitating the construction of cell graphs. Then, the neighboring nuclei will be clustering as clusters. These clusters serve as nodes in the cell-based graph, with node attributes determined by the standard deviation of nuclei sizes. Edges are established by calculating the geometric distance of the cluster centers based on their geometric coordinates with a distance threshold. Both patch- and cell-based approaches play integral roles in integrating histopathology and genomics data, particularly as more biological data become available. Graph-based models offer an efficient framework for capturing cross-modality differences. Beyond the graph neural network, graph transformers also show a promising capability to analyze non-euclidean graph structures for molecular biomarker prediction. Furthermore, a graph transformer architecture with local attention [35] is proposed for genomics profile alteration understanding via pathological image. In the proposed graph transformer architecture, DenseNet121 is used to extract features from pathological images, which are used as graph node attributions. Meanwhile, k-nearest neighbor graph edge construction is used to provide topological information among nodes. Such an operation allows the proposed method to explore the correlation between local and spatial morphology.

2.2 Pathology-and-genomics Multi-modal Analysis Enhances Survival Prediction

With the advancement of computational capabilities, multimodal datasets (e.g., imaging and genomic data) are increasingly integrated into disease diagnosis analysis. Multimodal data fusion can help with exploring the biologically cancer-related patterns and aggregate complementary information from various modalities to enhance cancer diagnosis and patient prognosis [36]. Several works successfully exploited the integration of image and non-image genomics biomarkers in assisting real-world clinical tasks, particularly in predicting patient survival outcomes. Typically, diverse domain-specific models are used to extract representations of serval modality data. These multimodal representations are then fused together for downstream tasks. Fusion strategies include several methods in various model training stages, such as multimodal feature concatenation or mapping multimodal features into a shared latent space via a well-designed optimization function. For example, a multimodal fusion framework [25] has been proposed to predict survival outcomes across various cancers by fusing histopathological images and genomic data, including mutations, CNV, and RNAseq data. This framework utilizes VGG19 to extract image-wise biomarker features and employs a cell-spatial graph to capture cellular associations. A Self-Normalizing Network (SNN) is used for extracting genomic features. Multimodal interactions are explored through Kronecker's product between unimodal features, with a gating-based attention mechanism controlling the relevance of each modality. PORPOISE, an extension of the previous multimodal framework [26], focuses on pan-cancer patient outcome understanding. It preserves the key components of the previous study [26] while excluding the cell-spatial graph representation. PORPOISE assigns learnable attention scores to image patches for each patient based on the contribution and importance of patient outcome prediction, enhancing feature representation. In contrast, the multimodal co-attention transformer (MCAT) framework [27] revolutionizes survival outcome prediction by incorporating image-genomic

interaction during intermediate model training stages rather than fusing them prior to a final prediction by using concatenation operations. In the tumor microenvironment, the genomics-guided co-attention strategy enables the capture and interpretation of complex genotype-phenotype interactions.

In addition to supervised approaches, an unsupervised method [24] can be used to reduce the human expert burden in data annotation. Such a method can leverage similarity evaluation among multimodal representations, which can project multimodal representations into a unified space during modality fusion. Deep highway networks [37] is used for extracting features on genomics data (e.g., gene expression and miRNA data), and SqueezeNet [38] is used for histopathological image feature extraction. Without model retraining, the multimodal feature representations can be achieved based on the previously trained model in the inference stage. The multimodal features are combined into a unified multimodal representation, facilitating overall survival prediction across pan-cancer datasets. The diverse combinations of multimodal data demonstrate varying performance, underscoring the potential of multimodal data in clinical diagnosis.

Different from natural vision-language datasets in the medical domain, the well-developed multimodal datasets remain limited in the medical domain. To address this challenge, a discrepancy and gradient-guided distillation framework [28] is developed based on a teacher-student knowledge distillation method to transfer pathology and genomic knowledge acquired by the teacher architecture to single modal (i.e., pathology-only) student model for glioma grading. Firstly, the teacher model integrates ResNet18-extracted pathological features with SNN-extracted genomic features. The multimodal knowledge fusion and refinement are achieved by training the model with a discrepancy-induced distillation loss. In the second stage, the student model distills multimodal information from the first-stage model and aggregates image-only knowledge from a mean-teacher model. Each modality of knowledge sends

a gradient on the student model, enabling it to be effectively employed in clinical-related tasks with only single-modality data while retaining multimodal knowledge.

CHAPTER 3: SPATIALLY-AWARE GRAPH NEURAL NETWORKS For CROSS-LEVEL MOLECULAR PROFILE ALTERATION PREDICTION IN COLON CANCER

3.1 Motivation

The advancements in graph convolutional networks (GCNs) have significantly advanced computational histopathology, particularly in terms of annotation efficiency and multi-scale context representation. Firstly, leveraging graph structures provides a feasible approach to represent entire slides in terms of tissue content connectivity. Such representations alleviate the need for fine-grained patch-wise label annotation, which is often time-intensive and impractical to encompass all ranges of tumor patches annotated by human experts. Secondly, graph structural representations enable the capture of multi-scale contexts by integrating global and local image-wise features, thereby enhancing disease outcome prediction. Thirdly, the utilization of graph structural representations facilitates interaction among spatially separated tiles, enabling a more flexible and comprehensive receptive field. These advancements mirror the workflow of human experts, who consider tumor environment, tissue contents, and their interactions rather than focusing solely on individual tumor tiles for diagnosing patient tissue status.

However, the high-resolution histopathological images do not naturally present a graph structure, so developing efficient graph representation is essential for model development and optimization. Current graph construction methodologies in histopathology can broadly be categorized into two approaches: patch-based and cell-based methods. Patch-based graph construction aims to extract information by considering the entire micro-environment, encompassing cells, and tissues, to capture comprehensive

tissue micro-environment and cell dynamics. In contrast, cell-based graph methods emphasize deriving possible biological insights from histopathology, exploring the relevance between different cells and tissue microenvironments using graph-based features [39]. It is worth noting that constructing a cell-based graph and conducting subsequent graph computations entail excessive computational complexity.

3.2 Methodology

3.2.1 Overview

In this dissertation, we proposed a spatially-aware graph neural network (GNN) architecture (see Figure 3.1)(g)) to predict cross-scale molecular profiles of gene mutations, copy number alterations and the level of protein expressions from histopathological images. We designed the image-to-graph transformation that converts the entire WSI into the spatially connected graph representation, where the spatial connections of tumor tiles are uniquely built upon the geometric coordinate from the raw WSI. The spatially-connectivity graph construction (Figure 3.1)(c) includes n tumoral tiles as graph nodes and their corresponding spatial grid coordinates. The graph nodes represent the identified WSI tiles and their attributes are the ResNet18-extracted image features. We also calculated the Euclidean distance between tile coordinates to determine the potential connectivity between tiles. Finally, we were able to generate spatially-connected graphs after graph node definition and edge connection. Multiple subgraphs were then constructed by node sampling with replacement on the WSI tiles to ensure a broad coverage of data samples. Next, our GNN-model architecture consists of five main modules, including a graph-based feature extractor, jumping knowledge structure, graph-level READOUT operation, multi-layer perceptron (MLP) classifier, and model ensemble strategy. We trained the proposed model on TCGA-COAD to predict the molecular outcome probabilities of the corresponding WSI slides. For each model, the input is a group of constructed spatially-connected subgraphs that are generated from each WSI slide. In the training and prediction

process, we use the constructed spatial subgraphs as the input of the proposed model to predict a series of molecular outcomes. For visualization and interpretation, we use four quantitative graph-structure measurement metrics on the constructed spatially-connected graphs ((Figure 3.1)(g)).

3.2.2 Spatially-connected Subgraph Construction

We designed the spatially-connected subgraph construction to represent the entire WSI by graph $G = (V, E)$, where $V = \{v_i, i \in N\}$ is the collection of graph vertices, $E = \{e_{ij}, i, j \in N\}$ is the collection of graph edges, and N is the number of vertices. We defined selected tiles as graph nodes in each subgraph. For each whole slide image (WSI), we only focused on analyzing tiles within the detected tumor region. We randomly selected a set of sampled patches $P = \{P_1, P_2, \cdots, P_N\}$ from all tumor tiles generated from WSIs, where N is the number of tiles. In statistics, random sampling is defined to facilitate generalization from the samples to the population, which ensures that sampling results approximate the population, especially when the entire population has been measured [40]. We measured all the tumor tissues in WSI, and thus, the multi-tile random sampling could maintain representative characteristics of the original WSI. Next, we utilized a ResNet18 feature extractor, which is pretrained on ImageNet, to extract tile features for each node as its feature matrix (e.g., node attributes, $X = \{x_i, i \in N\}$). The spatial distance between two tiles determines whether there exists a graph edge e_{ij} between two vertices (e.g., v_i and v_j). The entire structure of graph edge connectivity is represented by the adjacency matrix A, where we determine it by the Euclidean distance between tiles's geometric coordinates located in their raw WSI. We construct an edge between two nodes by calculating the Euclidean distance among them with a distance threshold. If the distance of the node is larger than the threshold, no edge is constructed between nodes. To determine the proper threshold, we calculated the mode value of from the statistical distribution of the spatial Euclidean distances among all pairs of tiles to determine the fixed thresh-

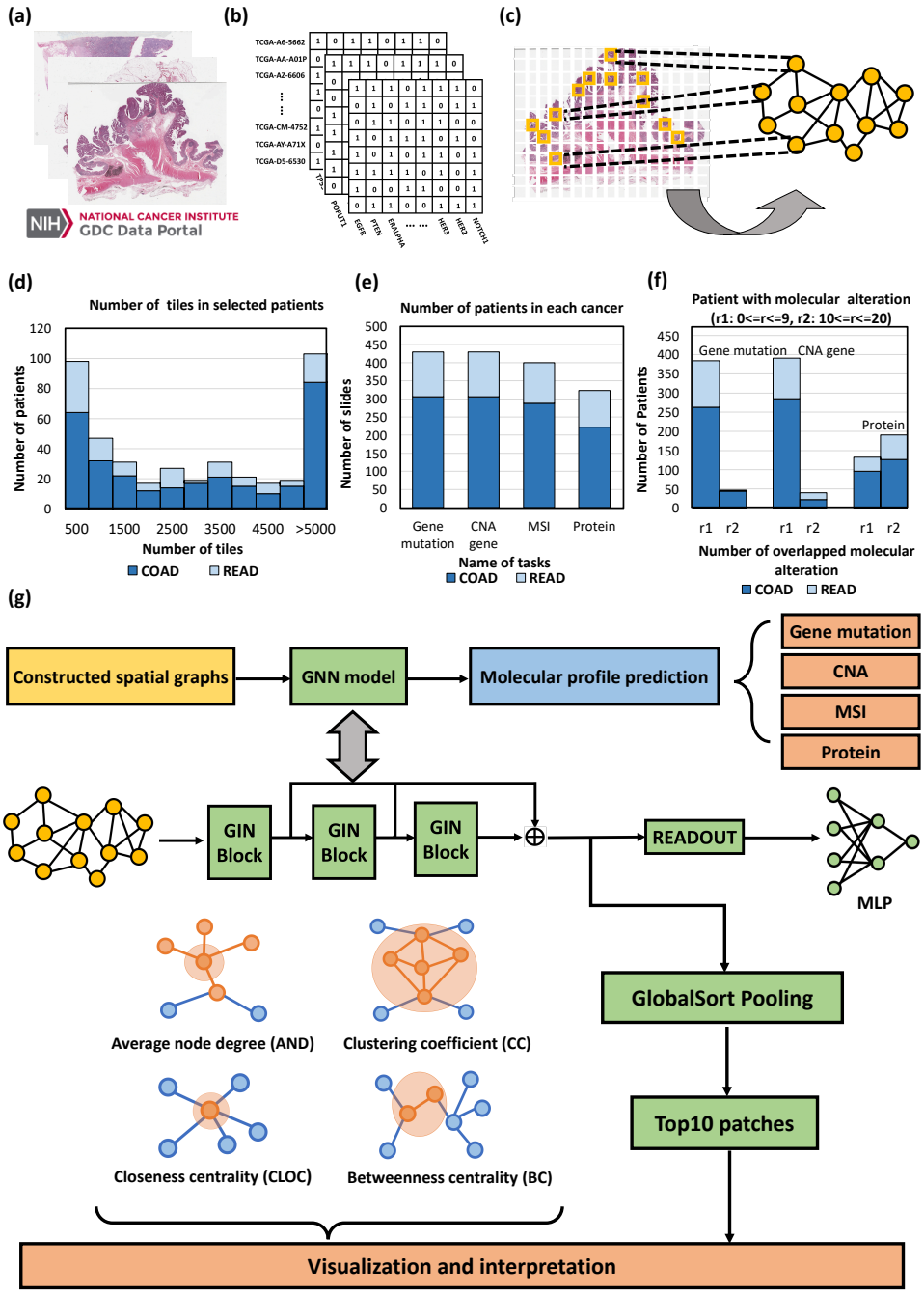


Figure 3.1: Pipeline and data statistics (reproduced with permission from Elsevier [1]). (a) High-resolution whole slide imaging data collection for patients. (b) Multi-scale molecular profile data collection including genetic mutation, copy number alteration, MSI status, and functional protein expressions. (c) Spatial graph construction of tumoral tiles. (d) Distribution of the number of tiles analyzed in individual patients. (e) Distribution of the number of tiles used in our dissertation for colon and rectum cancer. (f) Distribution of patients with overlapped molecular profiles for our integrative analysis. (g) The pipeline of the graph networks model training, prediction and interpretation process.

old value. Hereby, the graph $G = (V, E)$ denotes a graph with node feature vectors X and the adjacency matrix A. The constructed graphs are known as non-isomorphic graphs because of the different number of graph nodes and edge connections. Repeating the above strategy, we generated multiple subgraphs for each tumor WSI. The nodes in the different subgraphs can be overlapped because of the node selection strategy mentioned in the section on histopathology data preprocessing and image tile selection.

3.2.3 Graph-based Feature Extractor

The graph-based feature representation contains graph node attributes and their topological structures. We utilized the graph isomorphism network (GIN) layers in our study as the graph convolutional layer to aggregate and update the node representations (e.g., node features are extracted by ResNet18) [41]. In particular, the GIN layer utilizes a neighborhood aggregation strategy that updates the node representations by AGGREGATE and COMBINE operations iteratively that are widely used in spatial-based graph models [42, 41]. In Equation (3.1) and (3.2), for each node v, the AGGREGATE operation aggregates information from its neighboring nodes. In contrast, the COMBINE operation can integrate the representations of the center node (i.e., v) and its neighboring node to update node v's representation. After k iterations of aggregation and combination, the node representation is able to contain topological information within its k-hop connected neighboring nodes. The k^{th} layer spatial-based graph convolutional network can be represented as

$$h_{N(v)}^{(k)} = AGGREGATE_k(h_u^{(k-1)}, \forall u \in N(v)) \tag{3.1}$$

$$h_v^{(k)} = \sigma(W^k \cdot CONCAT(h_v^{k-1}, h_{N(v)}^{(k)}) \tag{3.2}$$

In GIN layers, the node representations are updated as follows:

$$h_v^{(k)} = LINEAR^{(k)}((1+\epsilon^{(k)}) \cdot h_v^{(k-1)} + \Sigma_{u \in N(v)} h_u^{(k-1)}) \tag{3.3}$$

The LINEAR is a single-layer perceptron that could represent the composition of functions. The ϵ is a learnable parameter or a fixed scale (by default with 0). An adjacency matrix of N×N is used for representing the structure of the input graph, and a node feature matrix of N×256. N is the number of constructed graph nodes (i.e., tumoral tiles) in our dissertation.

3.2.4 Jumping Knowledge Structure

Our WSI-based graph contains both dense and sparse connectivity nodes since each WSI has a complex spatial distribution of cancerous regions. To consider different levels of feature representations, we used the jumping knowledge (JK) structure to emphasize the integration of useful information obtained from all depths of network layers [43]. The JK structure aggregates node representation from each previous convolution layer to the last convolution layer by a max-pooling $max(h_v^{(1)}, \cdots, h_v^{(k)})$ to combine the node embeddings which are generated from each layer. By this design, the jumping knowledge connectivity between different convolutional layers could adaptively select the most evident representation from each layer. In other words, the model can select the most fitted neighborhood size for each node as needed in the training toward the proper node representation. Therefore, in our dissertation, the use of the JK structure allows spatial information integration from an adaptive range of nodes (i.e., tumoral tiles).

3.2.5 Graph-level READOUT Operation

Our focus is placed on the graph classification task that requires a function to convert the node embeddings into graph embedding. We thus used the GlobalAddPooling as a READOUT function to integrate node representations and produce a representation for each graph. As shown in Figure 3.1(g), we utilized jumping connectivity

to concatenate all layers. The graph representation can be written as:

$$h_G = READOUT(max(h_v^{(1)}, \cdots, h_v^{(k)}), v \in G) \tag{3.4}$$

k is 3 in our dissertation.

3.2.6 Multi-layer Perceptron (MLP) Classifier

We leveraged the MLP classifier to generate prediction results based on the graph-wise features extracted from the previous step. The MLP classifier consists of three fully connected (FC) layers with activation functions. We designed the first two fully connected layers to have 128 and 256 neurons, while the last FC layer has two neurons for the binary classification tasks.

3.2.7 Subgraph Model Ensemble Strategy

Our ensemble strategy utilized the majority vote to aggregate all subgraphs' prediction outcomes derived from the same WSI scan. To do so, we averaged the prediction scores from each subgraph model to achieve the slide-level prediction outcomes. The ensemble strategy is motivated by the fact that a high-resolution WSI can contain a large number of tumor tiles (e.g., 13k) that allows us to explore the diversity of WSI characteristics via combinations from individual subgraphs. We highlighted that ensemble learning allowed us to increase its generalization power by exploiting the advance of multiple spatial subgraphs for which the predictive error can be reduced by the majority vote.

3.2.8 Visualization and Graph Measurements

We evaluated the contribution of graph nodes by utilizing the global sort pooling [44]. Global sort pooling is devised to arrange the node features in descending order, determined by the outputs of their final feature channel. The outcome of the last convolutional layer is the first k nodes(e.g., the k is 10 as the top 10 tiles). The global sort pooling layer is used only for node contribution analysis. In other words,

the node contribution evaluation results of the global sort pooling layer will be used for model parameter training. We selected key tiles from the top ten nodes for each trained model and saved the tiles for visualization and interpretation. Figure 3.3 to Figure 3.5 provides the visualization of the top 10 tiles in each model and their located areas in the slides. For genetic mutation prediction, considering the clinical significance, we selected the mutated or CNA genes that are highly related to cancer treatment decisions. For protein expression prediction, we selected the proteins that are related to the evolution of colorectal cancer or the predictive mutated or copy number alteration gene in our previous experiments.

In addition, we utilized key graph measurements to provide an understanding of the constructed graph structures [45]. The average node degree calculates the average degree of the neighborhood of each node. We utilized average node degree to delineate the connectivity between nodes and their neighbors. The clustering coefficient quantifies the tendency of nodes in a graph to form clusters. Closeness centrality identifies nodes that have easy access to other nodes, with higher values indicating closer proximity to all other nodes. Betweenness centrality computes the total number of shortest paths between pairs of nodes and identifies nodes that serve as bridges connecting different parts of the network.

3.3 Experiments and Results

3.3.1 Image Data Collection and Selection

We used whole slide images of patients from The Cancer Genome Atlas Database and specifically focused on Colon Adenocarcinoma (TCGA-COAD) dataset and Rectum Adenocarcinoma (TCGA-READ) [46], which contain 459 Formalin-Fixed Paraffin-Embedded (FFPE) stained histopathology WSIs of colon tumor and 165 WSIs of rectum tumor. We collected the Clinical Proteomic Tumor Analysis Consortium (CPTAC) dataset, which contains 161 Fresh-Frozen (FF) WSIs of colon cancer tumors as the external validation dataset [47]. We show the statistics of patients in

Table 3.1: Patient characteristics from the collected TCGA-COAD cohort, TCGA-READ, and CPTAC-COAD cohort (reproduced with permission from Elsevier [1]).

	TCGA-COAD (n=306)	TCGA-READ (n=123)	CPTAC-COAD (n=94)
Age(year)			
Average	65.47	64.82	64.23
Sex, n(%)			
Male	156 (50.98)	66 (53.65)	39 (41.30)
Female	150 (49.02)	66 (46.35)	39 (58.70)
Stages, n(%)			
I/IA	48 (15.68)	16 (13.00)	9 (9.57)
II/IIA/IIB/IIC	112 (36.60)	44 (35.77)	34 (36.17)
III/IIIA/IIIB/IIIC	92 (30.07)	36 (29.27)	44 (46.81)
IV/IVA/IVB	45 (14.70)	19 (15.44)	7 (7.45)
N/A	9 (2.94)	8 (6.50)	0

Table 3.1.

We selected FFPE WSI slides according to the following criteria: (1) The slide exhibits blur-freed regions or tissue in regions with abnormal stains; (2) The slide presents sufficient and visible tumor regions; (3) One slide per patient comes with available gene mutation, copy number alteration (e.g., amplifications and deletions), microsatellite instability, and proteomic information. After preprocessing, we collected 306 patients with 40X magnification (0.25 microns/pixel) in TCGA-COAD. We selected the slide in mpp=0.25 microns/pixel due to its higher resolution to reflect image details than others. A similar selection approach is applied to the validation cohort of TCGA-READ and CPTAC-COAD. For TCGA-READ and CPTAC-COAD datasets, we collected 123 and 94 patients with WSI slides and associated molecular information. For microsatellite instability status (MSI) classification, we selected 288 slides in TCGA-COAD, 112 slides in TCGA-READ, and 94 slides in CPTAC-COAD with the available MSI records. For proteomics analysis, we obtained high-quality proteomic profiles generated by the antibody-based technique of reverse phase protein array (RPPA) from the TCPA database [13]. To ensure sufficient amounts of WSIs as training samples, we covered a wide range of available data samples in our

study. We selected the top 20 mutated genes in colon cancer with a mutation rate of at least 15% from patients, and for copy number alteration, we selected 20 genes with a mutation rate of at least 7.5%. For the validation tasks of rectum cancer, we selected 20 mutated genes in colon cancer, which were mutated at least 7% in the 123 patients, and for copy number alteration, we selected 20 mutated genes, with mutation percentages at least larger than 6%. For both colon cancer and rectum cancer, the cut-off ratio of high- and low-level protein expression was between 47% 52% based on the original TCPA proteomics records.3,4 For the external validation of colon cancer on CPTAC-COAD, we selected twenty genes in colon cancer with a mutation ratio of at least 12%. To achieve the CNA prediction, we determined five genes that have at least a 5% mutation rate. After WSI preprocessing (e.g., tile extraction and tumoral tile selection) and graph construction, a total of 670,901 tiles were used for the evaluation of colon cancer, and 225,146 tiles were used for the validation of rectum cancer.

3.3.2 Molecular Profile Collection and Label Identification

We identified the associated colorectal genetic mutational profiles and microsatellite instability status from Cbioportal (https://www.cbioportal.org/) [48]. Also, we collected protein expression profiles based on the reported clinical relevance of colon cancer and rectum cancer from The Cancer Proteome Atlas [13, 14]. Given the gene mutation rates, we focused on the top 20 frequently-mutated genes (e.g., APC, RYR1, KRAS, PIK3CA, TP53, TTN, SYNE1, OBSCN, FAT3, DNAH11, MUC16, FAT4, ZFHX4, LRP1B, FBXW7, CSMD1, RYR2, DNAH5, FLG, FAT3, DNAH11, CSMD3). We also collected the top 20 copy number alteration genes (e.g., CCSER1, COX4I2, CSMD1, DEFB118, DUSP15, FOXS1, ID1, MACROD2, MYLK2, HCK, KIF3B, PDRG1, PLAGL2, POFUT1, RBFOX1, REM1, TM9SF4, TPX2, TSPY26P, WWOX) that are associated with colorectal cancer evolution over various clinical stages. We obtained 20 functional protein expression profiles (e.g., ARID1A, BRAF,

P53, EGFR, STAT3_pY705, PTEN, EGFR_pY1173, HER3, SRC_pY416, BCL2, BRCA2, NOTCH1, CMYC, CMET_pY1235, ACC1, ACC_pS79, ATM, ERALPHA, HER2, AMPKALPHA_pT172) that have shown clinical relevance with targeted therapy [14]. For each type of mutational profile in each patient, we assigned the outcome label as a positive class if mutated and as a negative class if non-mutated. For microsatellite instability status classification, we assigned the status label as positive class if it is microsatellite instability, otherwise, we assigned it as negative class. For protein expression, the patient samples are separated into two groups by the median value of protein expression. If the value of protein expression is larger than the median value, we identified the sample with a high degree of protein expression and assigned it as a positive class. Otherwise, the sample was labeled with a degree of protein expression and assigned as a negative class.

3.3.3 Experimental Setting and Implementations

In our dissertation, we used Macenko's method for color normalization across all WSI slides to remove the bias of slide color. We split the WSI into non-overlapping square tiles with 512 pixels x 512 pixels edge length [49]. We utilized the OTSU segmentation algorithm to localize the histopathological tissue area (including tumor and non-tumor tissues) [50]. To further detect tumor regions, we first pretrained the ResNet on ImageNet and then fine-tuned the model (i.e., we retrained the entire model while modifying the output size of the last output layer as 2) on the NCT-CRC-HE-100K dataset [51], which contained 100,000 image tiles of colorectal cancer with pathologist-delineated, single-tissue regions as tumoral ground truth. As proved in a previous study [19], Resnet18 is more efficient for training and yields a better classification performance than models of Alexnet and VGG to reduce the risk of overfitting. We randomly spit the dataset into three sets: training set (70%), validation set (15%), and test set (15%) by following the previous study [19]. Also, for the input size of the tumor detection model, we resized WSI tiles to 224 x 224 pixels that

follow previous settings [19, 32]. We achieved >99% accuracy in the test set. The fine-tuned tumor detection model is applied to our WSI tissue data, and we found that the false-tolerant tumor region delineation is effective since the tile-focused analysis does not require precise tumor pixel segmentation. After tumor detection, we selected large amounts of tumor tiles from WSI for the tile-connected graph development, as detailed next. We chose to generate spatially-connected subgraphs via downsampling with replacement. For the slides containing a large tumor region, we randomly selected five subsets of tiles to build subgraphs. The number of selected tiles in each subset was set to 1,000, which is able to represent approximately 82% of the tissue area for the sake of computational efficiency. For the slides that only contain a small-sized tumor (e.g., the total number of tiles from the tumor region is smaller than 1,000), we kept all tumor tiles within the tumor for the subgraph construction. Together, the aggregated amount of non-overlapped tumor tiles from subgraphs of the entire tumor region and we found such sampling brings a good trade-off between model performance in prediction and computation efficiency.

To facilitate model training, we randomly duplicated the constructed graphs from the minority class in the training set for sample class balance. Data balancing ensures that the model learns characteristics equally from the majority and minority classes and avoids potential false positive or negative predictions. The benefit of duplication balancing class without requiring further data collection or augmentation. Notably, we only duplicated data samples in the training stage, and we always kept the real positive-and-negative ratio at the validation stage. The optimal hyperparameters were obtained by a grid search. We always kept the same hyperparameter settings for each prediction task to ensure that the differences only came from the variants of the model architectures. We set an initial learning rate of 1e-3 with Adam optimizer, where the batch size is 64. for training all models [52]. In our dissertation, we designed all predictions as classification tasks. Hence, we use the cross-entropy

as the loss function for model training and parameter optimization. In all prediction tasks (e.g., gene mutation, copy number alteration, MSI status, and protein expression prediction), we randomly initialized the weight parameters of the models before training. The model of each task was trained and evaluated separately without intervention. For the TCGA-COAD cohort, we used 10-fold cross-validation to train and evaluate the primary performance of our model. The entire TCGA-COAD dataset is randomly split into ten groups (e.g., ten-fold cross-validation) that have a similar number of samples. Each group is used as the evaluation set, which keeps the original positive and negative ratio of the dataset. When one group is used as an evaluation set, the class balancing is utilized for the other nine groups of data (e.g., training set) without changing the class ratio in the evaluation set. When one group is used as an evaluation set, the other nine groups of data are used for model training. Especially our data split (e.g., 10 folds) was slide-wise, which guarantees that the tiles in the same WSI will not appear simultaneously in different folds. The overall performance was reported based on the slide-level AUCs drawn for the concatenated results from all 10 folds. Next, we used all available slides from the TCGA-READ and CPTAC-COAD cohorts as the external validation data. In particular, we trained and selected a top model trained on the TCGA-COAD for one type of prediction task and directly validated on TCGA-READ and CPTAC-COAD for the same type of task without any transfer learning. The validation strategy is challenging due to the differences between two different datasets, such as their original collection source, screen instrument, and staining process. We used four Tesla V100 SXM2 GPUs for the prediction tasks.

To quantify the performance of our approach in molecular profile prediction tasks, we utilize the area under the curve (AUC) for model evaluation. The AUC represents the area under the ROC curve, which is plotted to represent the true positive rate against the false positive rate for different threshold values. AUC provides insight

into a model's ability to differentiate between classes. ROC curves and AUC are particularly effective for severely imbalanced classification problems with few samples of the minority class. For each AUC value, we calculate the 95% confidence interval (CI) using 1,000 bootstraps to determine the uncertainty of AUC. Additionally, we employ a student t-test to compute the significance by using a p-value (i.e., significance level set at $p < 0.05$).

3.3.4 Results

Cancer evolution is inherently associated with genetic alterations [53]. Detecting key genetic mutations is critical to assess staging, prognosis, and treatment for patients. In our dissertation, we evaluated the performance by the prediction scores of AUC, their 95% CIs, and student t-test p values. Our model achieved high-level performance on predicting multiple genetic mutations as shown in Figure 3.2(a), Figure 3.2(d) (top 10 predictable genes) and Table A.1 (the full results of predictable genes). In particular, we found that KRAS mutation (AUC 80.16) is well predicted by our approach (Table A.1), which has been recognized as a key determinant for measuring resistance to anti-EGFR therapy of colon cancer [54]. We also achieved a good prediction performance of TP53 (Table A.1) mutation (AUC 81.68 (95% CI 77.94-85.50)), a notable prognostic biomarker for colorectal patients treatment with 5FU chemotherapy [55]. As shown in Figure 3.2(a) and Figure 3.2(d), our image-based models well predict a panel outcome of gene mutations (top 10 predictable genes) with full prediction results in Table A.1.

The alternation of DNA fragments can cause the Copy number alterations (CNA), which are determined as the somatic change [56]. The accurate prediction of cCNA leads to the identification of relevant oncogenes, which is crucial for accurate diagnostics and therapy decision-making [57]. Following the same training process, as seen in Figure 3.2(b), Figure 3.2(d) (top 10 predictable genes) and Table A.2 (the full results of predictable genes), our model performed strongly (all AUC > 85.00)

in predicting the top 10 CNA genes in colon cancer. For instance, both POFUT1 (AUC 87.99 (95% CI 77.31-92.24)) and PLAG2 (AUC 90.55 (95% CI 86.02-94.89)) were highly predictive from our findings.

The signature of functional protein expression can be used to determine cancer progression, metastasis, and treatment that are not faithfully reflected by genetic alterations [13, 14]. Compared with genetic changes, protein-level activities occur at a functional level that is closely associated with cellular biology and drug development. We here present differential evidence for a comprehensive panel of key functional protein-expression degrees in colon cancer (see Figure 3.2(c)-(f) and Table A.3). For instance, the PTEN protein expression is predictable in our dissertation (AUC 86.01 (95% CI 81.97-90.06)), which represents a unique protein marker for predicting response to the treatment of Cetuximab [58]. The result of protein expression of HER3 (AUC 85.59 (95% CI 81.39-89.48)), broadens our positive findings since it is viewed as a determinant for poor prognosis of colon cancer [59].

To assess the cross-cancer generalization power of the model, we developed the model on the TCGA-COAD dataset while externally validating it on the rectum cohort from TCGA-READ without leveraging transfer learning. Multiple genetic mutations were confirmed predictive by our model as shown in Figure 3.2(g) and Figure 3.2(j) (e.g., top 10 predictable genes) and Table A.1 (e.g., the entire result of predictable genes). For instance, our model could predict KRAS mutation (the results is shown in Table A.1) on rectum cancer (AUC 71.02 (95% CI 63.39-89.48)) which is highly valuable to predict non-response to anti-EGFR target therapy (cetuximab and panitumumab) [54, 60]. Our model also achieved a high prediction performance on ZFHX4 (AUC 81.80 (95% CI 72.20-89.70)) that is associated with poor prognosis of patients. Additionally, we found potential predictive variables on the CNA status in rectum cancer with clinical relevance [12] for CNA prediction in Figure 3.2(h), Figure 3.2(k) (top 10 predictable genes), and Table A.2 (the rest of predictable CNA

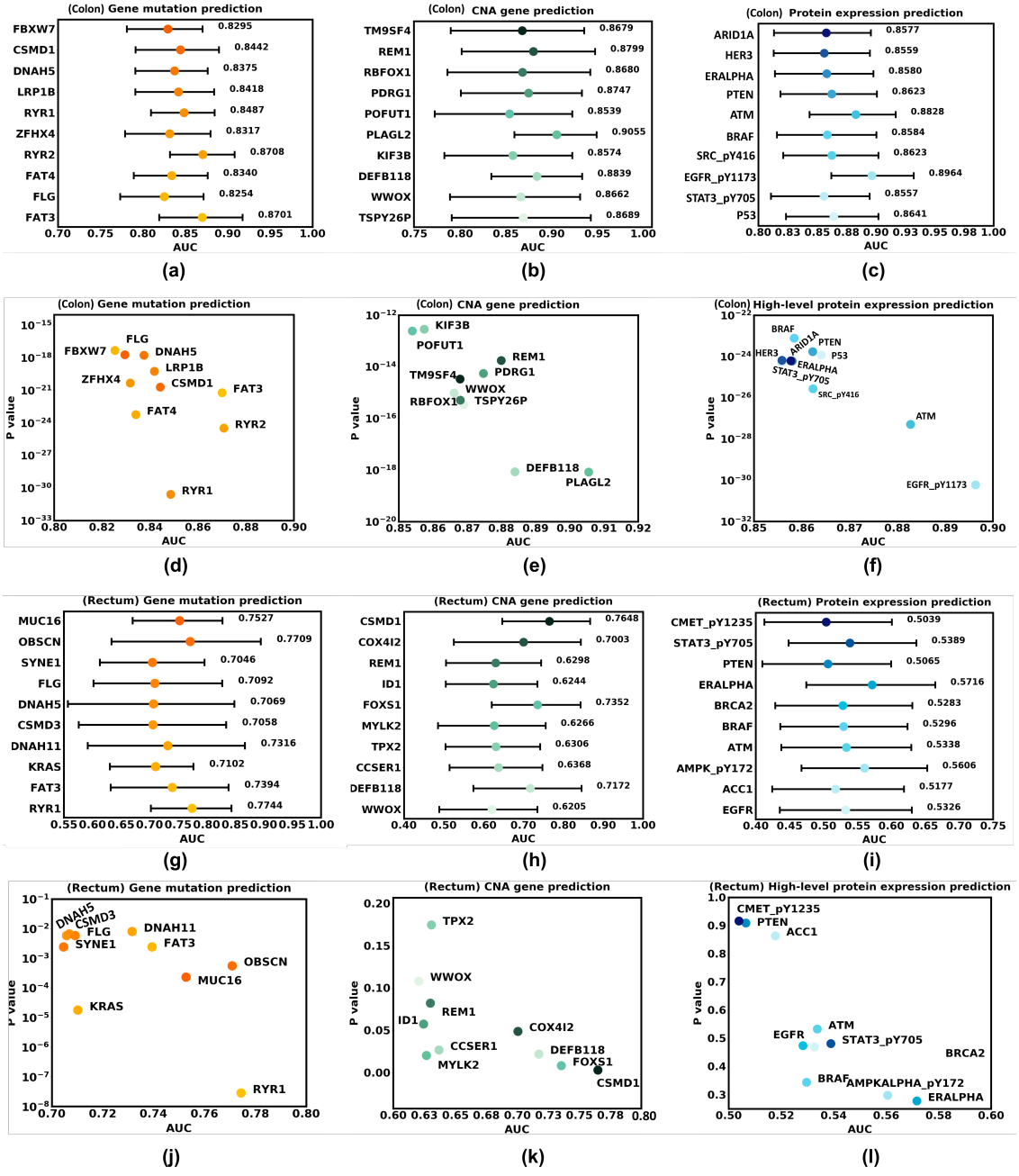


Figure 3.2: Molecular profile prediction results (reproduced with permission from Elsevier [1]). The GNN-based model was trained to predict the molecular profile outcomes (e.g., gene mutation, CNA gene, and protein expression) on colon cancer (TCGA-COAD) and validation results on rectum cancer (TCGA-READ). For each molecular profile, we show the prediction performance of AUC values with student t-test P value for the prediction scores (the significance level of 0.05). (a)–(c) The prediction results and its 95% CI in TCGA-COAD. (d)–(f) The prediction results and their P-values in TCGA-COAD. (g)–(i) The prediction results and their 95% CI in TCGA-READ cohort. (d)–(f) The prediction results and their P-values in TCGA-READ.

genes). We reported the level of protein expression prediction performance in Figure 3.2(i) and Figure 3.2(l) (top 10 predictable proteins) and Table A.3 (the rest of predictable proteins).

We further validate the model's potential generalization by training the model on the TCGA-COAD while directly validating it on the CPTAC-COAD with FF slides. We recognize positive findings on the CPTAC-COAD to inform the model's usefulness. For example, the model could validate DNAH5 (the results are shown in Table A.4) mutation (AUC 76.16 (95% CI 67.11-83.55)) that is highly associated with poor prognosis in colon cancer [61]. Also, we were able to predict FLG (the results are shown in Table A.4) mutation prediction (AUC 73.45 (95% CI 63.26-83.25)) on CPTAC-COAD, which is associated with loss of barrier function and deregulation of immune response [62]. Furthermore, multiple gene mutations confirm the prediction power of our model on the CPTAC-COAD dataset (Table A.4 and Table A.5).

We used our approach to achieve comparable performance (AUC 83.92 (95% CI 77.42-87.59)) of microsatellite instability status classification (MSI) in colon cancer. The validated finding is lower in rectum cancer (AUC 61.28 (95% CI 53.28-67.93)) due to the inherent cancer difference, meanwhile the additional MSI prediction evidence is positive (AUC 73.15 (95% CI 63.21-83.13)) on the CPTAC-COAD cohort despite of the slide format variance. The reliable imaging examination of MSI markers is ongoing, and our findings reiterate supportive evidence that predictive signals of MSI outcome were available [19].

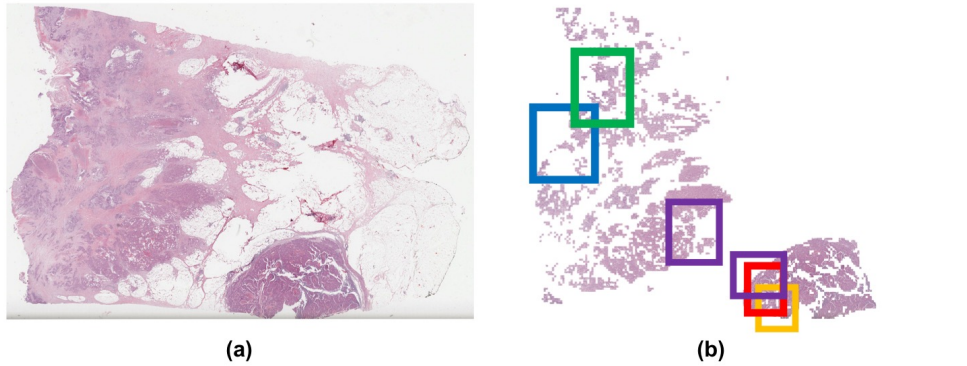
Despite the inherent gap between cancer types and image formats, we achieved a set of good findings. As seen in Table A.1 to Table A.3 and Table A.4 to Table A.5, we achieved positive gene mutation and CNA gene prediction results, such as ZFHX4 (AUC on TCGA-COAD is 83.17 (95% CI 78.00-87.98), AUC on TCGA-READ is 81.80 (95% CI 72.20-89.70)), CSMD1 (AUC on TCGA-COAD is 79.86 (95% CI 73.08-85.67), AUC on TCGA-READ is 76.48 (95% CI 64.78-86.71)). In addition, similar

findings include DNAH11 (AUC on TCGA-COAD is 82.42 (95% CI 77.16-87.75), AUC on CPTAC-COAD is 82.01 (95% CI 74.16-88.82)), and CCSER1 (AUC on and TCGA-COAD is 81.90 (95% CI 77.16-86.54), AUC on CPTAC-COAD is 78.50 (95% CI 67.87-87.34)). Interestingly, two molecular profiles could even be predicted better, such as CSMD3 (AUC on TCGA-COAD 82.17 (95% CI 77.82- 86.57), AUC on CPTAC-COAD 82.90 (95% CI 73.69-90.71)) and FOXS1 (AUC on TCGA-COAD 79.83 (95% CI 73.18-88.14), AUC on CPTAC-COAD 86.08 (95% CI 79.67-91.74)).

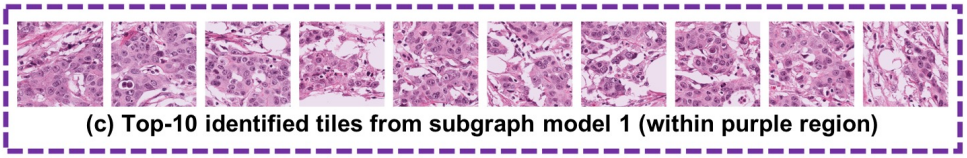
The lack of model understanding and interpretation of results has been a heightened concern for deep learning applications in medical domain research. Our graph network model employs a global sort pooling mechanism to provide possible interpretability. We display the top 10 tiles with the highest contribution based on the entire graph representation from each subgraph model (Figure 3.3-Figure 3.6). In Figure 3.3, we illustrated the result of TP53 mutation by the ensemble prediction from five subgraph models, which are separately trained by tile subgraphs generated from the entire WSI. Identified from each subgraph model, these top image tiles tend to be spatially distributed (colored regions), meaning that such a spatial characterization is of substantial interest for assessing molecular status in WSI. Also, the graph structure with a higher node degree and closeness centrality value (Figure 3.4-3.5 (h)) than the average statistics (Table 3.2) is informative by yielding accurate prediction for PLAGL2 copy number alteration (Figure 3.4) and PTEN protein expression (Figure 3.5).

3.3.5 Ablation Study

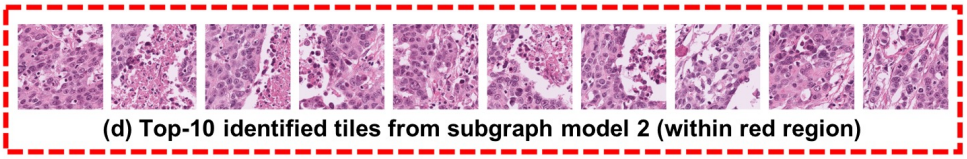
We designed ablation studies to analyze the performance of the proposed methods and make a comparison to baseline methods. We first designed a comparison between whether using ensemble strategy in the proposed workflow (Figure 3.7 (c)-(e)), which is a key factor for integrating multiple tile-connected graphs. Overall, the ensemble results were higher than those of other individual models without ensemble strategy,



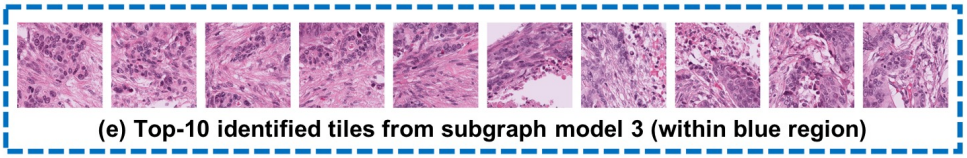
(a) (b)



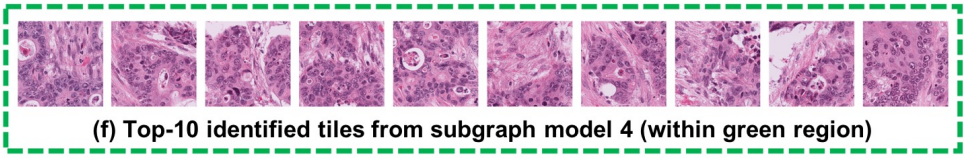
(c) Top-10 identified tiles from subgraph model 1 (within purple region)



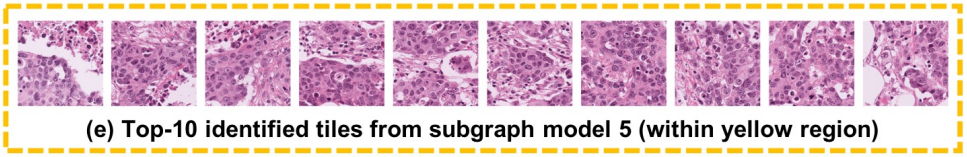
(d) Top-10 identified tiles from subgraph model 2 (within red region)



(e) Top-10 identified tiles from subgraph model 3 (within blue region)



(f) Top-10 identified tiles from subgraph model 4 (within green region)

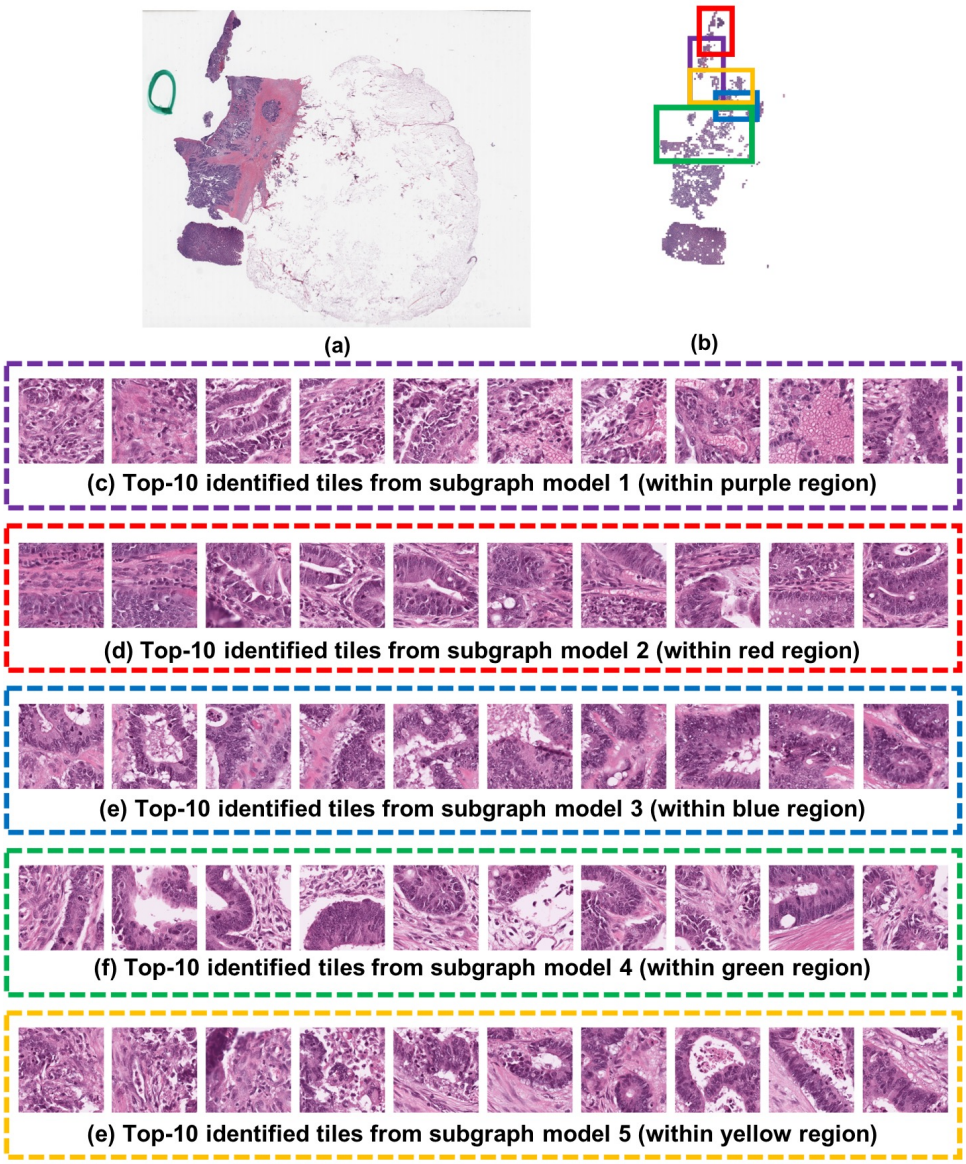


(e) Top-10 identified tiles from subgraph model 5 (within yellow region)

Statistics of subgraphs	Number of nodes	Node degree	Clustering coefficient	Closeness centrality	Betweenness centrality
	1000	688.5485	0.8610	0.7505	0.0004

(h)

Figure 3.3: TP53 mutation prediction on TCGA-COAD (reproduced with permission from Elsevier [1]). (a) The original WSI with TP53 mutation outcome. (b) Highlighted regions marked by the five subgraph models within the WSI. Different colors represent different key tile regions from subgraph models. (c)-(g) The zoom-in view of the identified top-10 tiles from five subgraph models which are ranked by their importance score in a decreasing order. From a pathologist's perspective, the gross necrosis is common in tiles from model 2 and model 3 while is rare in tiles from model1, model 4, and model 5. In addition, single cell necrosis is common in tiles from model 1 while are rare in model 5. (h) The average statistical results of the graph measurements among five subgraphs. Such graph measurements reflect the network structure of subgraphs.



Statistics of subgraphs	Number of node	Node degree	Clustering coefficient	Closeness centrality	Betweenness centrality
	1000	915.4283	0.9400	0.9177	0.0001

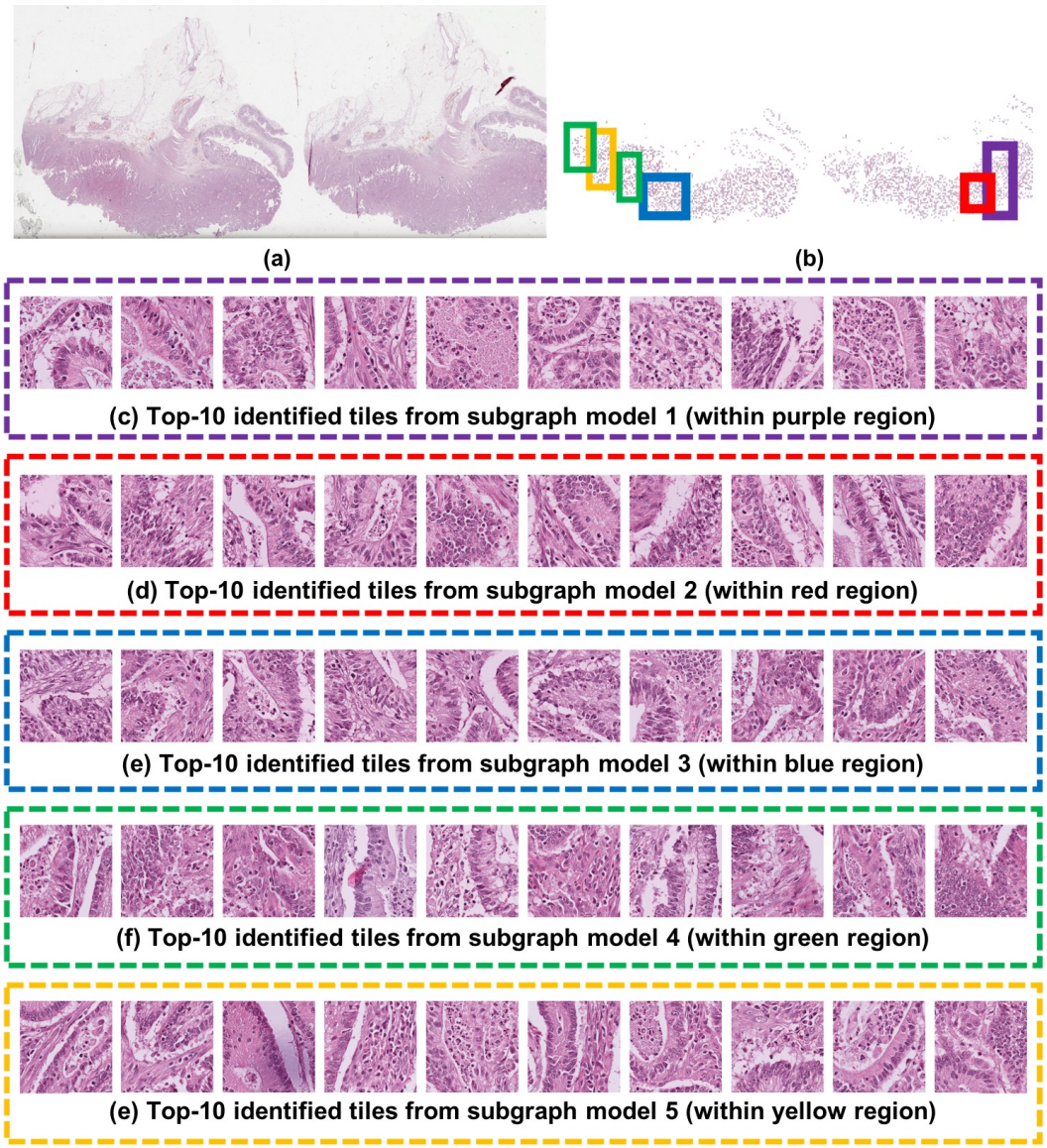
(h)

Figure 3.4: PLAGL2 CNA prediction on TCGA-COAD (reproduced with permission from Elsevier [1]). (a) The original WSI with PLAGL2 CNA. (b) Highlighted regions marked by the five subgraph models within the WSI. Different colors represent different key tile regions from subgraph models. (c)-(g) The zoom-in view of the identified top-10 tiles from five subgraph models which are ranked by their importance score in a decreasing order. Tiles from model 2 to model 4 almost do not contain lymphocytes while include rare apoptotic cells in model 2 and model 3. Furthermore, about 40% tiles from model 1 contain single cell necrosis. (h) The average statistical results of the graph measurements among five subgraphs. Such graph measurements reflect the network structure of subgraphs.

Statistics of subgraphs	Number of node	Node degree	Clustering coefficient	Closeness centrality	Betweenness centrality
	1000	759.7134	0.8750	0.7968	0.0003

(h)

Figure 3.5: PTEN protein expression prediction on TCGA-COAD (reproduced with permission from Elsevier [1]). (a) The original WSI with PTEN protein. (b) Highlighted regions marked by the five subgraph models within the WSI. Different colors represent different key tile regions from subgraph models. (c)-(g) The zoom-in view of the identified top-10 tiles from five subgraph models which are ranked by their importance score in a decreasing order. Tiles from model 1 to model 5 include mucinous tumor cells with background fibrosis. Significant surrounding lymphocytes are includes in model 2 to model 4. Furthermore, single cell necrosis is visible in model 5. (h) The average statistical results of the graph measurements among five subgraphs. Such graph measurements reflect the network structure of subgraphs.



Statistics of subgraphs	Number of node	Node degree	Clustering coefficient	Closeness centrality	Betweenness centrality
	1000	425.9053	0.8210	0.5223	0.0010

(h)

Figure 3.6: MSI status prediction on TCGA-COAD (reproduced with permission from Elsevier [1]). (a) The original WSI with MSI. (b) Highlighted regions marked by the five subgraph models within the WSI. Different colors represent different key tile regions from subgraph models. (c)-(g) The zoom-in view of the identified top-10 tiles from five subgraph models which are ranked by their importance score in a decreasing order. Tiles from model 1 to model 4 include mucinous tumor cells and tumor necrosis. In addition, tumor necrosis is common in tiles from model 2 to model 5, and mucinous tumor cells are common in tiles from model 2 to model 4. (h) The average statistical results of the graph measurements among five subgraphs. Such graph measurements reflect the network structure of subgraphs.

Table 3.2: The average statistics of graph measurements on TCGA-COAD and TCGA-READ, and CPTAC-COAD among all patients (reproduced with permission from Elsevier [1]).

	Number of nodes (SD)	Node degree (SD)	Clustering coefficient (SD)	Closeness centrality (SD)	Betweenness centrality (SD)
TCGA COAD	816.1025 (267.56)	600.1998 (328.1868)	0.9274 (0.1317)	0.7972 (0.2014)	0.0004 (0.0008)
TCGA READ	758.5920 (214.08)	564.7778 (342.6462)	0.9410 (0.1427)	0.8109 (0.2208)	0.0005 (0.0015)
CPTAC COAD	752.3900 (234.98)	668.2027 (308.4395)	0.9665 (0.0386)	0.9092 (0.1347)	0.0001 (0.0003)

and models without ensemble strategy displayed various results. Such variety could probably be explained by the intra-tumor heterogeneity that makes the performance of subgraphs different. Next, we ablated the aggregation strategy of jumping connectivity between convolutional layers in the GNN model. We leveraged three layer-wise aggregation strategies (e.g., max-pooling, concatenation, and LSTM-based attention aggregation) to integrate the node embeddings achieved by each previous convolutional layer [43]. As shown in Figure 3.7 (a), the max-pooling aggregation strategy achieves the best result, and LSTM-attention aggregation maintains a relatively stable performance. Furthermore, we assessed the distance threshold as a factor that determines the edge connectivity between graph nodes. A larger distance threshold leads to a denser connected graph, indicating that more graph edges are permitted to be connected. As seen in Figure 3.7 (b), the width and height of the original tiles are fixed at 512, and we set the distance threshold as their multiples (such as T = 25, 45, and 85 times). We found that dense graphs (T = 85) tend to be informative with higher results than other settings in genetic mutation, CNA gene, and protein expression degree prediction. Finally, we utilized the ResNet18 classifier as a baseline for the convolutional networks approach to compare with our GNN-based approach. To fairly compare our model with the ResNet18 model, we followed the exact same setting (e.g., data splitting and up-sampling augmentation during model

training). Further, we chose the same tiles for training ResNet18 without any data augmentation. The ResNet18 was pretrained on the ImageNet while we only trained the parameters in the last ten layers of the model and kept freezing weights in other layers, which followed the experiment setting in the previous study [19]. For genetic mutation, the convolutional approach has the average AUC of ResNet18 (AUC 64.87), which is evidently lower than the average AUC of our model (AUC 82.77) on the TCGA-COAD cohort. For MSI mutation, the AUC of ResNet18 (AUC 74.62) is also lower than the AUC of our model (AUC 83.92). Overall, we identified that the ensemble strategy of multiple subgraphs combined with jumping connectivity between convolutional layers achieved favorable results based on systematic comparisons.

To compare with the previous study [29], we conducted ablation experiments using gene mutations as benchmark targets on the TCGA-CRC-DX cohort (e.g., the combined TCGA-COAD and TCGA-READ FFPE diagnostic tissue slide cohort). We randomly split the entire TCGA-CDC-DX cohort into three groups (3-fold cross-validation) that have a similar number of samples for MSI status and gene mutation. Each group is used as the evaluation set in turn that keeps the original class ratio (e.g., the ratio between positive vs negative, and the ratio between TCGA-COAD and TCGA-READ patients) of the dataset. When one group is used as an evaluation set, the class balancing (e.g., only for positive and negative classes) is utilized for the other two groups of data (e.g., model training set) without changing the class ratio in the evaluation set. Especially our data split (e.g., 3 folds) was slide-wise, which guarantees that the tiles in the same WSI will not appear simultaneously in different folds. The mean cross-validated slide-level AUC values is used for performance evaluation. We found that our method outperformed a series set of results from the study [29], including APC (AUC 68.09 versus 65.40), PIK3CA (AUC 65.35 vs 62.20), and KRAS (AUC 64.46 vs 60.40) mutation prediction. Also, we achieved a comparable performance of TP53 (AUC 65.31 vs 68.50).

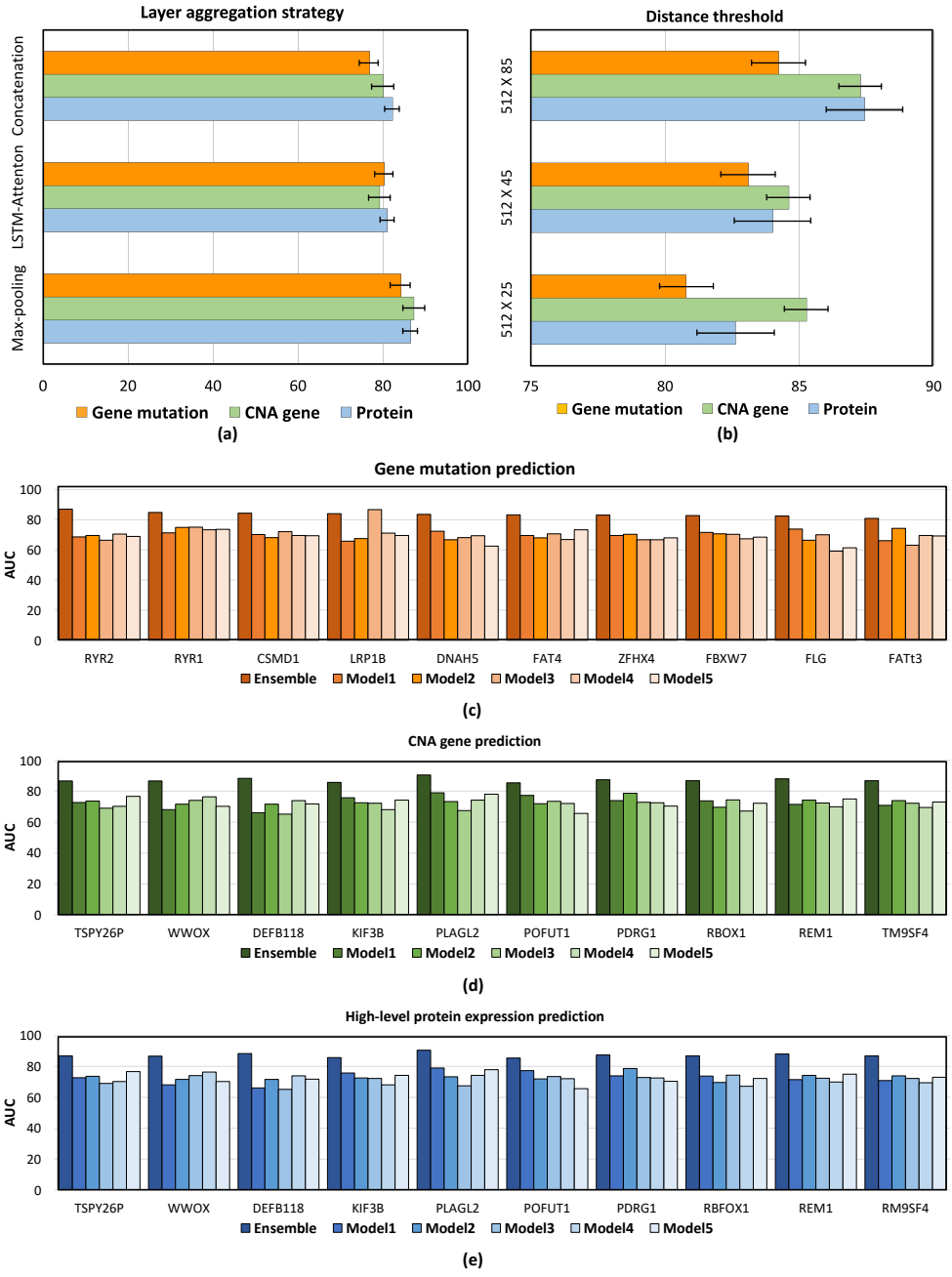


Figure 3.7: Ablation study of graph networks model performance (reproduced with permission from Elsevier [1]). (a) Layer aggregation strategy. We compared the average performance between different aggregation methods between GIN blocks, including max-pooling, LSTM-attention, and concatenation aggregation for three tasks, including gene mutation, CNA gene, and protein status prediction. Max-pooling is ranked the top result among all variations of blocks (b) Distance threshold. We averaged the AUC value of the top 10 predictions for each task. Outputs of distance threshold kept a relatively stable performance (all AUC>75%) with the choice of 512 x 85 as our desired selection. (c)-(f) Performance comparison between different subgraph models and ensemble model results. For the top 10 predicted genes, the performance between models with ensemble operation consistently outperformed all remaining models without ensemble operation.

3.4 Discussion

We proposed a graph neural network approach to explore spatial information via interactions of tumoral tiles of whole slide imaging (WSI). The presence of spatial and topological structures in histopathology is well documented but seldom explored in the context of quantitative cancer imaging and machine learning [63]. our dissertation emphasizes spatial context to construct tile-connected graphs to represent histopathological slides without explicit tile annotation, offering an efficient means to address intra-tumor spatial heterogeneity that is crucial to understand patient outcome in colon cancer [11]. In particular, our findings demonstrated that a broad range of molecular-histopathological associations was found to (i) infer prognostic value (e.g., KRAS and TP53 mutations), (ii) assess cell progression (e.g., PLAGL2 and POFUT1 copy number alterations), and (iii) identify targeted therapies (e.g., EGFR protein expressions) in colon cancer.

The rapid growth of whole-slide histopathology promises to uncover more meaningful genome-imaging associations via data integration [19]. Our analysis emphasizes a synergistic approach to the prediction and understanding of colon cancer based on molecular profiles in mutation, copy number alteration, and functional proteomics. In particular, proteomics exemplifies an emerging field to extend our landscape of genomic signature, which permits the direct discovery of diagnostic biomarkers from a cancer cellular perspective [64]. By definition, protein dynamics represent their own biological and cellular traits to complement roles of mRNA expressions [13, 14]. However, predictive analytics of proteomics profiles and their associations with other molecular signatures have not been explicitly researched in histopathology. In our dissertation, we achieved good predictions on both TP53 gene mutation prediction (AUC 81.68) and P53 protein expression prediction (AUC 86.41). From the perspective of cancer evolution, these findings reinforced our understanding that the symbolic TP53 mutation could promote colon cancer evolution leading to the abnormal protein

expression of P53 [55]. In addition, we identified the joint evidence that the protein expressions of Notch1 and copy number alterations of POFUT1 and PLAGL2 can be predicted because of their biological relationship [65]. our dissertation also achieved a good prediction of the functional protein BRAF (AUC 85.84 (95% CI 81.68-90.03)) and EGFR_pY1173 protein (AUC 89.64 (95% CI 86.29-93.19)), both of which are the parts of the EGFR-MAPK pathway to reflect the robustness of our dissertation. Thus our dissertation makes it possible to observe cross-scale molecular activities via histopathology that were not reported in previous studies. Also, diagnosis and therapy differ considerably between colon and rectum cancers, and our results offered helpful evidence that key mutational outcomes (e.g., ZFHX4 with AUC$>$ 80% and RYR1 with AUC$>$77% on both cancers) can be predicted to enhance the potential clinical utility of our approach.

The image-to-graph transformation in our dissertation opens up perspectives for analyzing tumoral spatial heterogeneity as seen in histopathology. Our contributions fall into multiple aspects, including spatial distance definition, image-tile graph construction and labeling, and topological interpretation of spatial characteristics. Driven by the observation that spatial heterogeneity is present within and across tumoral tiles in the entire cancer microenvironment, the proposed spatial distance builds upon tiles' physical geometric coordinates to objectively capture tumoral regional differences. In addition, our tile-based graph representation enables whole slide-level predictions, avoiding the uncertainty of tile label assignment for a particular molecular outcome. Such tile-based graph does not involve extra pre-preprocessing like nuclei or tissue segmentation which likely brings unfavorable performance variance.28 Assessing the full repertoire of multi-sized tiles is neither practical nor likely, given the excessive combinations of tiles required; thus, we focused on maximizing the information gleaned through efficient tile samplings. To faithfully depict the tile distribution, the whole-slide tile sampling creates an unbiased space allowing for the

subgraph construction from the divided tumoral tiles, which enhances model generalization and maintains a reasonable trade-off between efficiency and accuracy. We further provided a graph structure interpretation to quantitatively reveal the spatial interactions of image tiles. Finally, our graph approach is purely data-driven on the aggregated tumor tiles and does not rely on conventional morphological patterns that have been routinely assessed by pathologists. Consequently, it can serve as an augmented tool to diagnose suspicious malignancies and locate differential regions via the identified tumoral tiles in histopathology.

The multigenic complexity presents a daunting challenge for understanding the underlying mechanisms of colon cancer, motivating us to leverage the macroscopic view of histopathology via powerful graph networks. The strength of our approach relies on its capability to explore the relational context among complex graph entities that are beyond the scope of standard convolutional approaches [66]. Our analysis provides a comprehensive histopathological representation by extracting local (within tile) and topological (among tiles) information simultaneously, enabling a direct correlation measurement among regional tissues via importance ranking. The multi-parameter evaluation further reveals the stability of the proposed shallow graph neural networks across multiple prediction tasks. In addition, we acknowledge that there is a significant lack of consensus guidelines on the definition and utility of the tumoral image-based tiles. To address this challenge and enable detailed distribution analysis, the adopted random down-sampling with replacement ensures enough tiles to be selected for subgraph model development [40]. Our ensemble strategy further presents a simple yet effective means to merge the dynamics of tiles by aggregating the prediction results between different tile-connected subgraph models.

Although exploring the potential relationship between histopathology and molecular profiles is promising, more multi-site clinical verifications are necessary to add translational potential in the clinic and assist pathologists in gaining insights for

the identification of molecular signatures in colon cancer and management of other cancers. Emerging techniques in spatial transcriptomics may provide highly defined annotations to locate fine-grained histopathological regions and further enhance deep-learning performance [67]. Also, we recognize that the class imbalance of molecular profiles is commonly seen across cancers, making the training samples insufficient to optimize the model development. For example, copy number alteration genes like TM9SF4, TPX2, TSPY26P, and WWOX only have about 7.69% mutation ratio in the cohort, although they represent meaningful clinical relevance in colon molecular pathology [68]. We recognize that data format differences of histopathology can impact model robustness for certain mutational outcome predictions. It is also meaningful to extend our graph analysis into the pan-cancer setting by assessing the model consistency across cancer types. Considering the limited number of data samples, we have not analyzed the joint molecular activity prediction that could provide knowledge about measuring complex image-genome relationships. The landscape of molecular, pathological, and predictive studies of cancer is changing rapidly, and the continued investigation of modeling long-tail characteristics of molecular classes will be crucial to uncover additional insights into genome-pathology associations in cancer.

3.5 Summary

In conclusion, we contributed a spatially-aware graph neural networks approach to predict molecular profiles and converted the WSI slides into graph structures with spatial-preserving information. Despite multiple levels of molecular heterogeneity, our findings offered a panel of predictable molecular profiles, including mutational outcomes, copy number alteration outcomes, MSI status, and protein expression from WSIs in colon cancer. Our computational approach provides a unique means to characterize spatial heterogeneity of colon cancer that has the potential generalization to uncover widespread imaging-molecular correlations, which impacts treatment determination, prognosis assessment, and improved management of colon cancer.

CHAPTER 4: PATHOLOGY-AND-GENOMICS MULTIMODAL TRANSFORMER FOR SURVIVAL OUTCOME PREDICTION

4.1 Motivation

We introduce a multimodal framework named **PathOmics** for survival outcome prediction by integrating the pathological and genomics characteristics (Figure 4.1). We showed our three contributions as below. **(1) Unsupervised multimodal data fusion:** Our unsupervised pretraining leverages the inherent interaction between morphological and molecular biomarkers (Figure 4.1a). To address the modality heterogeneity gap between images and genomics data, we map the multimodal embeddings into a shared latent space by assessing their relevance. Notably, the pretrained model employs relevance-guided modality fusion to extract cross-modal patterns. **(2) Flexible modality finetuning:** A significant contribution of our multimodal framework is that it can utilize the benefits from both unsupervised pretraining and supervised finetuning data fusion (Figure 4.1b). Consequently, task-specific finetuning extends dataset utility (Figure 4.1b and c), which allows flexible data modality usage (e.g., both single- and multi-modal data). **(3) Data efficiency with limited data size:** Our approach achieves comparable performance even with fewer finetuned data (e.g., utilizing only 50% of the finetuned data) compared to using the entire finetuning dataset.

4.2 Methodology

4.2.1 Overview

Figure 4.1 illustrates our multimodal transformer framework. Our method includes unsupervised multimodal data fusion pretraining and supervised flexible-modal fine-

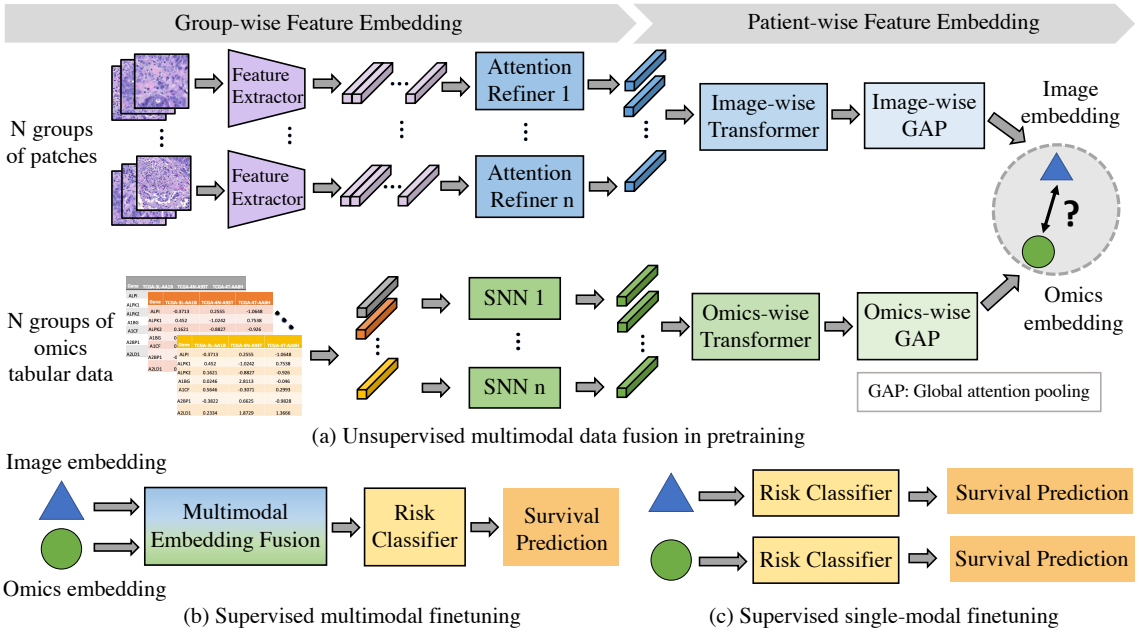


Figure 4.1: Workflow overview of the pathology-and-genomics multimodal transformer (**PathOmics**) for survival prediction (reproduced with permission from Springer Nature [2]). In (a), we show the pipeline of extracting image and genomics feature embedding via an unsupervised pretraining towards multimodal data fusion. In (b) and (c), our supervised finetuning scheme could flexibly handle multiple types of data for prognostic prediction. With the multimodal pretrained model backbones, both multi- or single-modal data can be applicable for our model finetuning

tuning. From Figure 4.1a, in the pretraining, our unsupervised data fusion aims to capture the interaction pattern of image and genomics features. Overall, we formulate the objective of multimodal feature learning by converting image patches and tabular genomics data into group-wise embeddings and then extracting multimodal patient-wise embeddings. More specifically, we construct group-wise representations for both image and genomics modalities. For image feature representation, we randomly divide image patches into groups. Meanwhile, for each type of genomics data, we construct groups of genes depending on their clinical relevance [69]. Next, as seen in Figure 4.1b and c, our approach enables three types of finetuning modal modes (i.e., multimodal, image-only, and genomics-only) towards prognostic prediction, expanding the downstream data utility from the pretrained model.

4.2.2 Group-wise Image and Genomics Embedding

We define the group-wise genomics representation by referring to $N = 8$ major functional groups obtained from [69]. Each group contains a list of well-defined molecular features related to cancer biology, including transcription factors, tumor suppression, cytokines and growth factors, cell differentiation markers, homeodomain proteins, translocated cancer genes, and protein kinases. The group-wise genomics representation is defined as $G_n \in \mathbb{R}^{1\times d_g}$, where $n \in N$, d_g is the attribute dimension in each group which could be various. To better extract high-dimensional group-wise genomics representation, we use a Self-Normalizing Network (SNN) together with scaled exponential linear units (SeLU) and Alpha Dropout for feature extraction to generate the group-wise embedding $G_n \in \mathbb{R}^{1\times 256}$ for each group.

For group-wise WSI representation, we first cropped all tissue-region image tiles from the entire WSI and extracted CNN-based (e.g., ResNet50) d_i-dimensional features for each image tile k as $h_k \in \mathbb{R}^{1\times d_i}$, where $d_i = 1,024$, $k \in K$ and K is the number of image patches. We construct the group-wise WSIs representation by randomly splitting image tile features into N groups (i.e., the same number as genomics categories). Therefore, group-wise image representation could be defined as $I_n \in \mathbb{R}^{k_n\times 1024}$, where $n \in N$ and k_n represents tile k in group n. Then we apply an attention-based refiner (ABR) [70], which is able to weight the feature embeddings in the group, together with a dimension deduction (e.g., fully-connected layers) to achieve the group-wise embedding. The ABR and the group-wise embedding $I_n \in \mathbb{R}^{1\times 256}$ are defined as:

$$a_k = \frac{epx\{w^T(tanh(V_1h_k) \odot (sigm(V_2h_k))\}}{\sum_{j=1}^{K} epx\{w^T(tanh(V_1h_j) \odot (sigm(V_2h_j))\}} \tag{4.1}$$

where w,V1 and V2 are the learnable parameters.

$$I_n = \sum_{k=1}^{K} a_k h_k \tag{4.2}$$

4.2.3 Patient-wise Multimodal Feature Embedding

To aggregate patient-wise multimodal feature embedding from the group-wise representations, as shown in Figure 4.1a, we propose a pathology-and-genomics multimodal model containing two model streams, including a pathological image and a genomics data stream. In each stream, we use the same architecture with different weights, which is updated separately in each modality stream. In the pathological image stream, the patient-wise image representation is aggregated by N group representations as $I_p \in \mathbb{R}^{N\times 256}$, where $p \in P$ and P is the number of patients. Similarly, the patient-wise genomics representation is aggregated as $G_p \in \mathbb{R}^{N\times 256}$. After generating patient-wise representation, we utilize two transformer layers [71] to extract feature embeddings for each modality as follows:

$$H_p^l = MSA(H_p) \tag{4.3}$$

where MSA refers to Multi-head Self-attention [71], l denotes layer index of the transformer, and H_p could either be I_p or G_p. MSA is the combination of k self-attention (SA) operations via concatenation operation. The SA uses d_k-dim patient embedding as the query Q, key K, and value V to learn paired relationship $a_{ij} \in A$ among $q_i \in Q$ and $k_i \in K$:

$$softmax(\frac{QK^T}{\sqrt{d_k}}) = A \tag{4.4}$$

$$SA(Q, K, V) = AV \tag{4.5}$$

Then, we construct global attention poolings [70] as Equation 4.4 to adaptively

determine a scored sum of each modality feature embeddings to finally construct patient-wise embedding as $I^p_{embedding} \in \mathbb{R}^{1\times256}$ and $G^p_{embedding} \in \mathbb{R}^{1\times256}$ in each modality.

4.2.4 Multimodal Fusion in Pretraining and Finetuning

Due to the domain gap between image and molecular feature heterogeneity, a proper design of multimodal fusion is crucial to advance integrative analysis. In the pretraining stage, we develop an unsupervised data fusion strategy by decreasing the mean square error (MSE) loss to map images and genomics embeddings into the same space. Ideally, the image and genomics embeddings belonging to the same patient should have a higher relevance to each other. MSE measures the average squared difference between multimodal embeddings. Sequentially, in the latent space, the pretrained model is developed to project the paired image and genomics embeddings to be closer, leading to novel insights into multimodal interaction.

$$\mathcal{L}_{fusion} = argmin\frac{1}{P}\sum_{p=1}^{P}((I^p_{embedding} - G^p_{embedding})^2) \tag{4.6}$$

In the single modality finetuning, even if we use image-only data, the model is able to produce genomic-related image feature embedding due to the multimodal knowledge aggregation already obtained from the model pretraining. As a result, our cross-modal information aggregation relaxes the modality requirement in the finetuning stage. As shown in Figure 4.1b, for multimodal finetuning, we deploy a concatenation layer to obtain the fused multimodal feature representation and implement a risk classifier (FC layer) to achieve the final survival stratification. As for single-modality finetuning mode in Figure 4.1c, we simply feed $I^p_{embedding}$ or $G^p_{embedding}$ into risk classifier for the final prognosis prediction. During the finetuning, we update the model parameters using a log-likelihood loss for the discrete-time survival model training [27]. We extend the definition and detailed proof of "discrete-time survival prediction" as follows. The

continuous event time $T_{j,continue} \in [t_r, t_r + 1)$ could be discretized as T_j, which is equal to r, where $r \in \{0, 1, 2, 3\}$ and j is the index of four non-overlapped intervals. The discrete ground truth is $Y_j \in \{0, 1, 2, 3\}$. With patient-wise embedding h_{final_j}, we define the hazard function $f_{hazard}(r|h_{final_j})$ as $P(T_j = r|T_j \geq r, h_{final_j})$, which is used for calculating the survival function (i.e., C-index calculation) $f_{surv}(r|h_{final_j})$ through $P(T_j > r|h_{final_j})$ (i.e., $\prod_{u=1}^{r}(1 - f_{hazard}(u|h_{final_j}))$). During the supervised finetuning, the log-likelihood loss for model parameter updation is defined as $-c_j \cdot log(f_{surv}(Y_j|h_{final_j})) - (1-c_j) \cdot log(f_{surv}(Y_j - 1|h_{final_j})) - (1-c_j) \cdot log(f_{hazard}(Y_j|h_{final_j}))$, where $c_j = 0$ means patient passed away during T_j and $c_j = 1$ means patient lived after T_j.

4.3 Experiments and Results

4.3.1 Datasets

All image and genomics data are publicly available. We collected WSIs from The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) dataset (CC-BY-3.0) [72, 47] and Rectum Adenocarcinoma (TCGA-READ) dataset (CC-BY-3.0) [73, 47], which contain 440 and 153 patients. We cropped each WSI into 512 $\times$ 512 non-overlapped patches. We also collected the corresponding tabular genomics data (e.g., mRNA sequence, copy number alteration, and methylation) with overall survival (OS) times and censorship statuses from Cbioportal [74, 48]. We removed the samples without the corresponding genomics data or ground truth of survival outcomes. Finally, we included 426 patients of TCGA-COAD and 145 patients of TCGA-READ.

4.3.2 Experimental Settings and Implementations

We implement two types of settings that involve internal and external datasets for model pretraining and finetuning. As shown in Figure 4.2a, we pretrain and finetune the model on the same dataset (i.e., internal setting). We split TCGA-

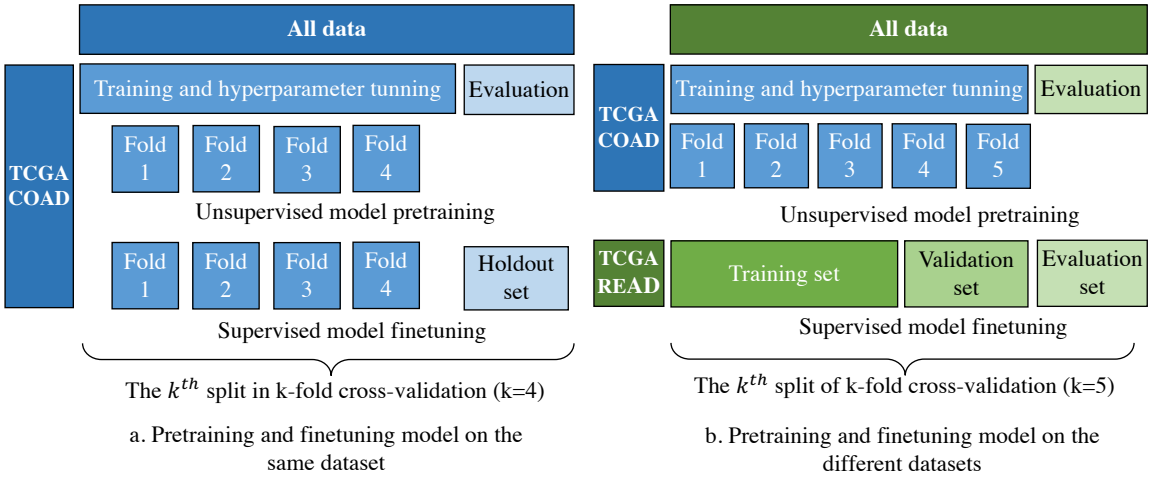


Figure 4.2: Dataset usage (reproduced with permission from Springer Nature [2]). In a, we use TCGA-COAD dataset for model pretraining, finetuning, and evaluation. In b, we use TCGA-COAD dataset for model pretraining. Then, we use TCGA-READ dataset to finetune and evaluate the pretrained models

COAD into training (80%) and holdout testing set (20%). Then, on the training set, we implement four-fold cross-validation for pretraining and hyperparameter-tuning in the finetuning. The test set is only used to evaluate the best finetuned models from each cross-validation split. For the external setting, we implement pretraining and finetuning on the different datasets, as shown in Figure 4.2b; we use TCGA-COAD for pretraining; Then, we only use TCGA-READ for finetuning and final evaluation. We use a five-fold cross-validation for pretraining while we only use the best pretrained models for finetuning. We split TCGA-READ into finetuning (60%), validation (20%), and evaluation set (20%). For all experiments, we calculate the average performance on the evaluation set across the best models.

The number of epochs for pretraining and finetuning is 25, and we set the batch size as 1; the learning rate is 1e-4 for pretraining and 5e-5 for finetuning with an Adam optimizer [52]. We used one 32GB Tesla V100 SXM2 GPU and Pytorch. The concordance index (C-index) is used to measure the survival prediction performance. We followed the previous studies [27, 25, 26] to partition the overall survival (OS) months into four non-overlapping intervals by using the quartiles of event times of

uncensored patients for discretized-survival C-index calculation (see Appendix 2). For each experiment, we reported the average C-index among three-times repeated experiments. Conceptionally, our method shares a similar idea to multiple instance learning (MIL) [75, 76]. Therefore, we include two types of baseline models, including the MIL-based models (DeepSet [77], AB-MIL [70], and TransMIL [78]) and MIL multimodal-based models (MCAT [27], PORPOISE [26]). We follow the same data split and processing, as well as the identical training hyperparameters and supervised fusion as above. Notably, there is no need for supervised finetuning for the baselines when using TCGA-COAD (Table **??**), because the supervised pretraining is already applied to the training set.

4.3.3 Results

In Table 4.1 and Table 4.2, our approach shows improved survival prediction performance on both TCGA-COAD and TCGA-READ datasets. Compared with supervised baselines, our unsupervised data fusion is able to extract the phenotype-genotype interaction features, leading to achieving a flexible finetuning for different data settings. With the multimodal pretraining and finetuning, our method outperforms state-of-the-art models by about 2% on TCGA-COAD and 4% TCGA-READ. We recognize that the combination of image and mRNA sequencing data leads to reflecting distinguishing survival outcomes. Remarkably, our model achieved positive results even using a single-modal finetuning when compared with baselines (more results in Appendix 3.1). In the meantime, on the TCGA-READ, our single-modality finetuned model achieves a better performance than multimodal finetuned baseline models (e.g., with model pretraining via image and methylation data, we have only used the image data for finetuning and achieved a C-index of 74.85%, which is about 4% higher than the best baseline models). We show that with a single-modal finetuning strategy, the model could generate meaningful embedding to combine image- and genomic-related patterns. In addition, our model reflects its efficiency on the limited

Table 4.1: The comparison of C-index performance on TCGA-COAD dataset. "Methy" is used as the abbreviation of Methylation

Model	Pretrain data modality	TCGA-COAD	
		Finetune data modality	C-index (STD)
DeepSets [77]	image+mRNA	-	58.70 (1.10)
	image+CNA	-	51.50 (2.60)
	image+Methy	-	65.61 (1.86)
AB-MIL [70]	image+mRNA	-	54.12 (2.88)
	image+CNA	-	54.68 (2.44)
	image+Methy	-	49.66 (1.58)
TransMIL [78]	image+mRNA	-	54.15 (1.02)
	image+CNA	-	59.80 (0.98)
	image+Methy	-	53.35 (1.78)
MCAT [27]	image+mRNA	-	65.02 (3.10)
	image+CNA	-	64.66 (2.31)
	image+Methy	-	60.98 (2.43)
PORPOISE [26]	image+mRNA	-	65.31 (1.26)
	image+CNA	-	57.32 (1.78)
	image+Methy	-	61.84 (1.10)
Ours	image+mRNA	image+mRNA	**67.32** (**1.69**)
		image	63.78 (1.22)
		mRNA	60.76 (0.88)
	image+CNA	image+CNA	61.19 (1.03)
		image	58.06 (1.54)
		CNA	56.43 (1.02)
	image+Methy	image+Methy	67.22 (1.67)
		image	60.43 (0.72)
		Methy	61.06 (1.34)

finetuning data (e.g., 75 patients are used for finetuning on TCGA-READ, which are only 22% of TCGA-COAD finetuning data). In Table 4.1 and Table 4.2, our method could yield better performance compared with baselines on the small dataset across the combination of images and multiple types of genomics data.

Table 4.2: The comparison of C-index performance on TCGA-READ dataset. "Methy" is used as the abbreviation of Methylation

Model	Pretrain data modality	TCGA-READ	
		Finetune data modality	C-index (STD)
DeepSets [77]	image+mRNA	image+mRNA	70.19 (1.45)
	image+CNA	image+CNA	62.50 (2.52)
	image+Methy	image+Methy	55.78 (1.22)
AB-MIL [70]	image+mRNA	image+mRNA	68.79 (1.44)
	image+CNA	image+CNA	66.72 (0.81)
	image+Methy	image+Methy	55.78 (1.22)
TransMIL [78]	image+mRNA	image+mRNA	67.91 (2.35)
	image+CNA	image+CNA	62.75 (1.92)
	image+Methy	image+Methy	53.09 (1.46)
MCAT [27]	image+mRNA	image+mRNA	70.27 (2.75)
	image+CNA	image+CNA	60.50 (1.25)
	image+Methy	image+Methy	59.78 (1.20)
PORPOI -SE [26]	image+mRNA	image+mRNA	68.18 (1.62)
	image+CNA	image+CNA	60.19 (1.48)
	image+Methy	image+Methy	68.80 (0.92)
Ours	image+mRNA	image+mRNA	74.35 (1.15)
		image	**74.85** (**0.37**)
		mRNA	59.61 (1.37)
	image+CNA	image+CNA	73.95 (1.05)
		image	71.18 (1.39)
		CNA	63.95 (0.55)
	image+Methy	image+Methy	71.80 (2.03)
		image	64.42 (0.72)
		Methy	65.42 (0.91)

4.3.4 Ablation Analysis

We verify the model efficiency by using fewer amounts of finetuning data in finetuning. For TCGA-COAD dataset, we include 50%, 25%, and 10% of the finetuning data. For the TCGA-READ dataset, as the number of uncensored patients is limited, we

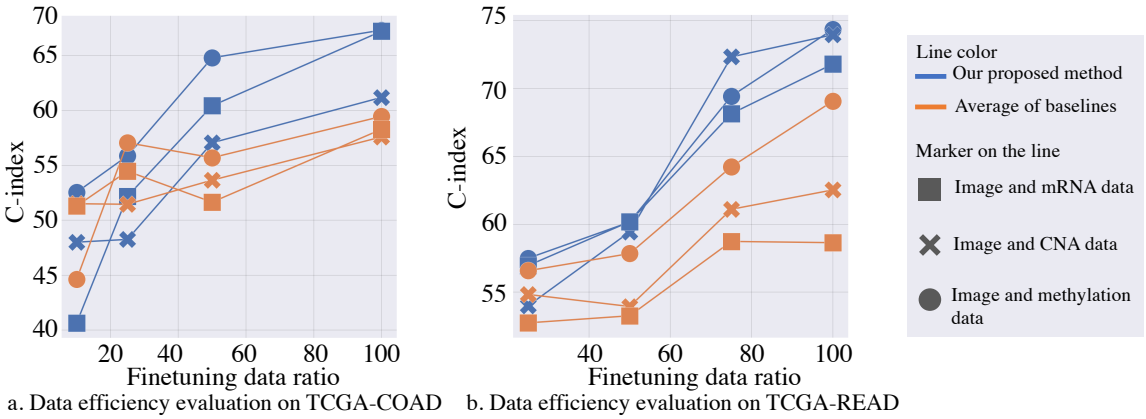


Figure 4.3: Ablation study (reproduced with permission from Springer Nature [2]). In (a) and (b), we evaluate the model efficiency by using fewer data for model finetuning on TCGA-COAD and TCGA-READ. We show the average C-index of baselines, the detailed results are shown in the Appendix 3.2

use 75%, 50%, and 25% of the finetuning data to allow at least one uncensored patient to be included for finetuning. As shown in Figure 4.3a, by using 50% of TCGA-COAD finetuning data, our approach achieves the C-index of 64.80%, which is higher than the average performance of baselines in several modalities. Similarly, in Figure 4.3b, our model retains a good performance by using 50% or 75% of TCGA-READ finetuning data compared with the average of C-index across baselines (e.g., 72.32% versus 64.23%). For evaluating the effect of cross-modality information extraction in the pretraining, we kept supervised model training (i.e., the finetuning stage) while removing the unsupervised pretraining. The performance is lower 2%-10% than ours on multi- and single-modality data. For evaluating the genomics data usage, we designed two settings: (1) combining all types of genomics data and categorizing them by groups; (2) removing category information while keeping using different types of genomics data separately. Our approach outperforms the above ablation studies by 3%-7% on TCGA-READ and performs similarly on TCGA-COAD. In addition, we replaced our unsupervised loss with cosine similarity loss; our approach outperforms the setting of using cosine similarity loss by 3%-6%.

4.4 Summary

Developing data-efficient multimodal learning is meaningful in advancing the patient survival assessment in a variety of clinical data scenarios. We demonstrated that the proposed PathOmics framework is useful for improving the survival prediction of colon and rectum cancer patients. Importantly, our approach opens up perspectives for exploring the key insights of intrinsic genotype-phenotype interactions in complex cancer data across modalities. Our finetuning approach broadens the scope of dataset inclusion, particularly for model finetuning and evaluation, while enhancing model efficiency on analyzing multimodal clinical data in real-world settings. In addition, the use of synthetic data and developing a foundation model training will be helpful to improve the robustness of multimodal data fusion [79, 80].

CHAPTER 5: CONTRASTIVE PATHOLOGY-AND-GENOMICS MULTIMODAL LEARNING FOR SURVIVAL OUTCOME PREDICTION

5.1 Motivation

Inspired by the previous success of PathOmics [2] in patient survival outcome prediction, image-genomics multimodal analysis becomes a viable solution enabling precise patient prognosis in real-world applications. Yet, several challenges remain to be addressed for better performance and more reliable multimodal knowledge acquisition. The pertaining scheme in PathOmics can potentially mix the unique characteristics among individual patients because of the missing scheme for distinguishing patients. To overcome this challenge, we propose an effective contrastive pathology and genomics pretraining scheme to capture the multimodal interactions while distinguishing the differences among different patients.

5.2 Preliminary

5.2.1 Multimodal Contrastive Pretraining

The advent of contrastive learning introduces a discriminative method aiming to lead similar samples to be closer to each other while making the different samples far from each other [81]. For instance, within the image domain, contrastive learning trains image encoders by generating augmented image data, optimizing to maximize the similarity between their projected embeddings while maximizing the dissimilarity with embeddings of other samples [82]. Notably, implementations like SimCLR [83], BYOL [84], and MOCO [85] have popularized contrastive learning, showing their promising feature representation capabilities across single modality studies, including images or text.

An effective contrastive loss function is widely used, which is called InfoNCE [86]:

$$L_{q,k+,k-} = -\log \frac{\exp(q \cdot k_+/\tau)}{\exp(q \cdot k_+/\tau) + \sum_{k_-} \exp(q \cdot k_-/\tau)}. \tag{5.1}$$

q denotes a query, k_+ represents the representation of a positively related (similar) key sample, and k_- is the representation of negatively related (dissimilar) key samples [87]. The parameter τ refers to a temperature hyper-parameter. In the pretext task of instance discrimination [88], a query and a key form a positive pair if they originate from the same image source. Conversely, if the query and key are from different sources, they constitute a negative pair. In an end-to-end setup, keys associated with negative pairs are sampled from the same batch and updated through backpropagation. SimCLR [83] employs the above mechanism, requiring a large number of batch sizes to provide a sizable set of negative pairs. Conversely, in the MoCo mechanism [85], the negative pairs are stored in a queue, while positive pairs of the queries and keys can be encoded during each training step.

Inspired by such latent representation learning ability, contrastive learning becomes a viable solution for multimodal information aggregation and embedding generation. Such as CLIP [84], which integrates images and language, is a leading model for multimodal analysis. It also shows promising generalizability by yielding good performance on zero-shot inference tasks. Such models are generally training on a large-scale web-curated dataset with millions of parameters, which are known as foundational models, such as UniCL [89], Florence [90], and ALIGN [91]. In the medical domain, there are limited studies in analyzing multimodal data by foundation models because of the limited size of medical datasets for model development. Recently, image and genomics data have been explored for multimodal knowledge aggregating, aiming to enhance disease diagnosis and prognosis performance (e.g., imaging and genetic data multimodal analysis [92]). Furthermore, self-supervised tabular and imaging models [93] have shown the possibility of utilizing multimodal models for clinical-related analysis,

Yet they use only two or four clinical features.

Among the introduced promising multimodal contrastive learning studies, CLIP is widely used several fields, including the medical domain. The proposed pair-wise pre-training strategy enables promising generalizability of downstream tasks. In detail, CLIP (Contrastive Language-Image Pretraining) [84] is a pretraining method developed by OpenAI, designed to fill the gap between images and texts. CLIP jointly optimizes a vision encoder and a text encoder to produce single-modality embedding. It ensures that image-text pairs are closely aligned while unpaired image and text are far from each other in a shared latent space. Unlike the previous methods that rely on extensive manual efforts or complex model architecture, CLIP introduces an efficient and simple means for multimodal information aggregation and yields promising performance, especially on zero-shot tasks.

5.2.2 Contrastive Pre-training Architecture and Principles

The architecture of CLIP integrates a vision encoder model with a language encoder model by following a loss function principle similar to that of InfoNCE. The visual encoder can be based on either ResNet [94] or Vision Transformer (ViT) [95], while the text encoder is selected as a transformer-based architecture like BERT [96]. During the pretraining stage, CLIP is fed by a batch of images and their corresponding text captions as input in each iteration. Similar to the single-modal encoding process, each single-modal embedding in CLIP is normalized and projected to a joint image-text latent space. The original images and texts are encoded into $I \in \mathbb{R}^{N \times D}$ and $T \in \mathbb{R}^{N \times D}$, respectively, where N denotes batch size, and D represents embedding dimensionality.

In CLIP, contrastive pretraining plays a crucial role in image-text modality knowledge alignment. Different from the conventional models that are supervised by a single or predefined task, CLIP learns the inherent interactions between paired image-text information by contrastive pretraining. In detail, N^2 image and text pairs can be

constructed by the batch size of N, where N matched pairs of image and text data (i.e., positive pairs) and $(N^2 - N)$ unmatched image-text pairs (i.e., negative pairs). The pretraining loss function for the image encoder is hence denoted as

$$L_{\text{img}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\Phi(I_i, T_i)/\tau)}{\sum_{j=1}^{N}\exp(\Phi(I_i, T_j)/\tau)}, \tag{5.2}$$

where $\Phi(\cdot,\cdot)$ is denoted as cosine similarity calculation, the symbol τ is a learnable temperature parameter, I_i and T_i refers to the ith image embedding and text embedding, respectively. The loss function for the text encoder is:

$$L_{\text{txt}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\Phi(T_i, I_i)/\tau)}{\sum_{j=1}^{N}\exp(\Phi(T_i, I_j)/\tau)}. \tag{5.3}$$

The total optimization loss function of CLIP is designed as the average of Equation 5.2 and Equation 5.2:

$$L_{\text{total}} = \frac{L_{\text{img}} + L_{\text{txt}}}{2}. \tag{5.4}$$

5.2.3 The Generalizability of CLIP on Zero-shot Tasks

With the advent of pretraining CLIP by predicting whether an image matches a textual caption instead of specific supervised tasks, CLIP is naturally suitable for applying to zero-shot scenarios. In the inference stage, CLIP will not be fine-tuned as conventional methods. Alternatively, the pretrained CLIP is used to generate the embedding of query images or text. Then, the generated image or text embeddings will be used to compare with the other embeddings. CLIP presents a unique means for solving zero-shot classification tasks as follows. I_1 represents the image features extracted by the image encoder for a query image x, and $\{W_i\}_{i=1}^{K}$ is a set of class embeddings generated by the text encoder of CLIP, where K denotes the number of classes, and each W_i is a text prompt resembling "a photo of a [CLASS]". The

probability of class prediction is calculated by:

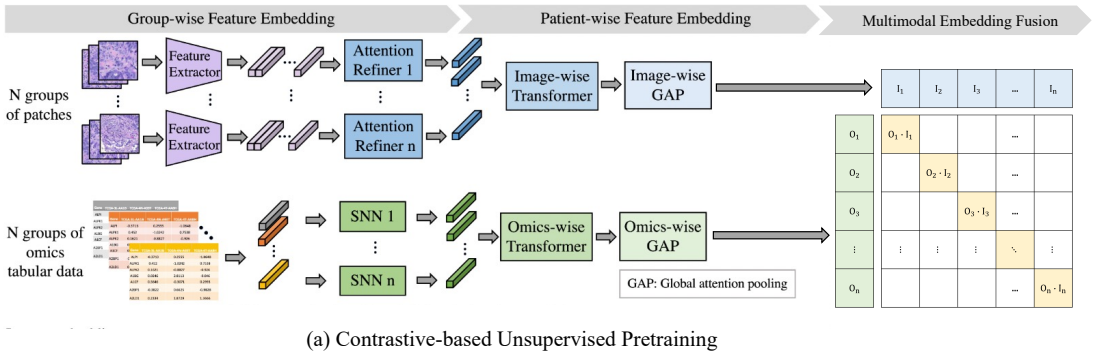
$$p(y = i|I_1) = \frac{\exp(\Phi(I_1, W_i)/\tau)}{\sum_{j=1}^{K} \exp(\Phi(I_1, W_j)/\tau)}, \tag{5.5}$$

where τ is a temperature parameter achieved during pretraining, and $\Phi(\cdot, \cdot)$ is the operation of the cosine similarity. Although originally trained on web-curated images and the corresponding text captions, CLIP has demonstrated its promising capability in several downstream tasks. The unique inference capability in zero-shot allows CLIP to understand a query image without explicit prior training.

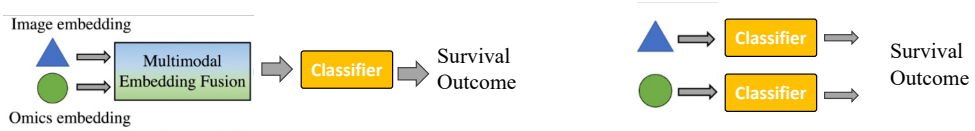
5.3 Methodology

5.3.1 Overview

We show our contrastive-based pathology-and-genomics multimodal framework (**C-PathOmics**) in Figure 5.1. Our framework develops an unsupervised pretraining approach for multimodal data fusion by exploiting contrastive learning, along with a supervised flexible-modal finetuning that allows for specific task alignment. In the unsupervised pretraining phase, illustrated in Figure 5.1a, our contrastive-based data fusion aims to capture the interaction pattern of pathological images and genomics embeddings while enhancing the distinction among different patients. We start by randomly partitioning image patches into groups [2]. Simultaneously, for each type of genomics data, we create groups of genes based on their clinical relevance [69]. Subsequently, we transform image patches and tabular genomics data into group-wise embeddings, followed by the integration of single-modal group-wise embeddings into multimodal patient-wise embeddings. In Figure 5.1b, our approach facilitates a modality-flexible finetuning strategy (e.g., multimodal, image-only, and genomics-only data), enabling enhanced patient prognostic prediction by broadening the downstream data utility derived from the pretrained model. Such a modality-flexible fine-tuning scheme also provides a viable solution for addressing real-world



(a) Contrastive-based Unsupervised Pretraining



(b) Multi-modal Multi-task Finetuning

Figure 5.1: Workflow overview of the contrastive-based pathology-and-genomics multimodal model (**C-PathOmics**) for survival prediction. In (a), we illustrate the pipeline for extracting image and genomics feature embeddings via contrastive-based unsupervised pretraining, facilitating multimodal data fusion. In (b), our modality-flexible supervised finetuning scheme can handle multiple data modalities for patient outcome prediction. Leveraging the multimodal pretrained model backbones, both multi- and single-modal data can be utilized for our model finetuning.

patient data modality missing concerns.

5.3.2 Contrastive Pathology-and-genomics Pretraining

To fill the domain gap between histopathological image and molecular characteristics heterogeneity, proper multimodal knowledge fusion is crucial in advancing integrative analysis and enabling precise patient outcome prediction. In the pretraining stage, different from the previous study [2], we develop an unsupervised data fusion strategy based on the success of contrastive multimodal pretraining. We hypothesized that the histopathological image and genomics embeddings belonging to the same patient should be more relevant to each other. Alternatively, the pathology-genomics pairs belonging to different patients should have a lower relevance to each other.

To bridge the gap between the heterogeneity between multimodal characteristics (i.e., image and genomics data), effective multimodal knowledge fusion is promising

for enabling precise patient outcome prediction. In the pretraining stage, unlike the approach in the previous study [2], we introduce an unsupervised pretraining strategy based on contrastive multimodal data fusion. Our hypothesis is that the histopathological images and genomics embeddings belonging to the same patient should have higher relevance compared to those belonging to different patients. We aim to develop a contrastive-based multimodal pretraining scheme for the model by mapping the paired image and genomics embeddings to be more relevant to the latent space while the unpaired pathology-genomics embeddings should be far from each other. Such contrastive-based multimodal knowledge integration is able to enhance the relevance among multiple modalities while distinguishing the differences between various patients. We list our algorithm for contrastive pathology-and-genomics pretraining as follows:

Input: P[n, h, w, c] - minibatch of pathological images

G[n, l] - minibatch of the corresponding genomic sequence of the images

$W_p[d_p, d_e]$ - learned projection of image for embedding

$W_g[d_g, d_e]$ - learned projection of text for embedding

t - learnable temperature parameter

Output: loss

`// Extract feature representations of each modality`

$P_f \leftarrow$ pathological_encoder(P) ; `// [`m`,` d_p`]`

$G_f \leftarrow$ genomics_encoder(G) ; `// [`n`,` d_g`]`

`// Joint multimodal embedding`

$P_e \leftarrow$ l2_normalize(np.dot(P_f, W_p), axis=1) ; `// [`m`,` d_e`]`

$G_e \leftarrow$ l2_normalize(np.dot(G_f, W_g), axis=1) ; `// [`n`,` d_e`]`

`// Scaled pairwise cosine similarities`

logits $\leftarrow$ np.dot($P_e, G_e^\top$) $\times$ exp(t) ; `// [`m`,` n`]`

`// Symmetric loss function`

labels $\leftarrow$ np.arange(n)

loss_p $\leftarrow$ cross_entropy(logits, labels, axis=0)

loss_g $\leftarrow$ cross_entropy(logits, labels, axis=1)

loss $\leftarrow$ (loss_p + loss_G)/2

return *loss*

Algorithm 1: Multimodal Symmetric Loss Algorithm

5.3.3 Modality-flexible Finetuning

Following the previous study [2], we use a modality-flexible finetuning strategy to extend the usage of our method. In the real-world scenario, such a modality-flexible finetuning strategy is able to introduce benefits to patients with missing data. The input in the finetuning stage can either be multimodal data (e.g., pathological

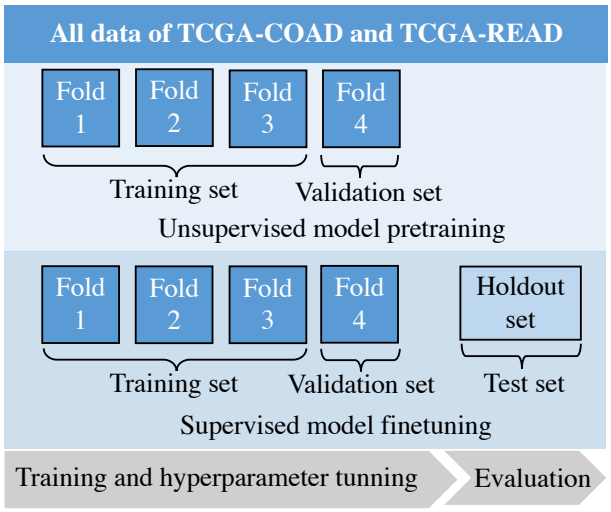


Figure 5.2: Dataset usage. we combine TCGA-COAD and TCGA-READ datasets for model pretraining, finetuning, and evaluation. In the figure, we use a four-fold cross-validation for model training and selection (i.e., validation) in both the unsupervised pretraining and supervised finetuning stages. In the evaluation stage, we use a hold-out test set for model performance evaluation to avoid data leakage issues and ensure the fairness of the evaluation.

image and genomics data) or single-modal data (e.g., image- or genomics-only data). In the stage of multimodal finetuning, we utilize a simple concatenation operation to integrate the feature representations from multiple modalities and exploit a risk classifier (FC layers) to predict patient survival outcomes. Regarding the single-modality finetuning scheme, we directly use either the image or genomics embedding as the input for the risk classifier to access the patient prognosis prediction. In finetuning, we update the parameters of the model using a log-likelihood loss for the discrete-time survival model training as introduced in[2].

5.4 Experiments and Results

5.4.1 Experimental Settings and Implementations

To include as much data as we can, we use both TCGA-COAD and TCGA-READ datasets (e.g., named TCGA-CRC) for further experiments. As shown in Figure 5.2, we pretrain and finetune the model on the TCGA-CRC dataset. We split TCGA-

CRC into a training set (80%) and a holdout testing set (20%). Subsequently, we perform four-fold cross-validation on the training set for both pretraining and finetuning. In K^{th} cross-validation split, where $k = 4$, we randomly select one fold as the validation fold for hyperparameter-tuning while training the model on the rest of the three folds. The hold-out test set is solely set to evaluate the best finetuned models, which are selected in each cross-validation split. The number of epochs is set to 25 and the batch size of 8 in both pretraining and finetuning. We set the initial learning rate as 5e-4 for pretraining and 1e-4 for finetuning. We also use a learning rate scheduler to decrease the learning rate in every ten epochs. We utilize the Adam optimizer [52] in our experiments. We did the experiments on a single 48GB Nvidia RTX A6000 GPU using Pytorch. Similar to the previous study [2], we use the concordance index (C-index) to measure survival outcome prediction performance by using the risk score, event time, and censored status. Following previous studies [27, 25, 26], we partition the overall survival (OS) months into four non-overlapping intervals by using the quartiles of event times of uncensored patients for discretized-survival C-index calculation. For each experiment, we report the average C-index among three times repeated experiments with different random seeds (i.e., 42, 1024, and 2048). Our method shares similarities with the concept of multiple instance learning (MIL) [75, 76]. Hence, we first contain MIL-based models (DeepSet [77], AB-MIL [70], and TransMIL [78]) as baseline models. Then, we include SOTA multimodal methods (MCAT [27], PORPOISE [26], and PathOmics [2]). We use the same dataset split, model training hyperparameters, and supervised fusion as mentioned above. Notably, supervised finetuning is not required for the baselines (Table 5.1), as they undergo supervised training directly on the training set without an unsupervised training stage.

Table 5.1: The comparison of C-index performance on TCGA-CRC dataset. "Methy" is used as the abbreviation of Methylation

Model	Pretrain data modality	TCGA-CRC	
		Finetune data modality	C-index (STD)
DeepSets [77]	image+mRNA	-	57.19 (4.04)
	image+CNA	-	56.38 (4.81)
	image+Methy	-	56.78 (2.70)
AB-MIL [70]	image+mRNA	-	60.80 (1.50)
	image+CNA	-	51.62 (1.51)
	image+Methy	-	58.49 (4.75)
TransMIL [78]	image+mRNA	-	60.72 (4.01)
	image+CNA	-	58.89 (3.06)
	image+Methy	-	52.47 (3.49)
MCAT [27]	image+mRNA	-	64.52 (3.79)
	image+CNA	-	56.22 (2.39)
	image+Methy	-	56.25 (2.75)
PORPOISE [26]	image+mRNA	-	57.43 (3.30)
	image+CNA	-	56.87 (1.66)
	image+Methy	-	54.59 (1.86)
PathOmics [2]	image+mRNA	image+mRNA	61.80 (1.43)
	image+CNA	image+CNA	64.53 (1.66)
	image+Methy	image+Methy	58.39 (2.11)
Ours	image+mRNA	image+mRNA	**66.03 (2.39)**
		image	56.22 (2.36)
		mRNA	**67.29 (2.72)**
	image+CNA	image+CNA	61.20 (2.19)
		image	62.14 (1.70)
		CNA	61.30 (3.48)
	image+Methy	image+Methy	59.06 (2.80)
		image	58.45 (1.55)
		Methy	57.06 (1.79)

5.4.2 Results

As shown in Table 5.1, our proposed method shows improved survival prediction performance on TCGA-CRC among different types of baseline models. Compared with supervised baselines, our contrastive-based unsupervised data fusion shows a promising capability in extracting the phenotype-genotype interaction features, leading to a good survival outcome prediction performance with flexible finetuning among different data modality settings. The state-of-the-art supervised baseline is MCAT [27], which yields a c-index of 64.52% for survival outcome prediction using multimodal data (i.e., image and miRNA data). Our proposed method is able to outperform 2% performance improvement compared with MCAT by using image and miRNA data for model pretraining and finetuning. Promisingly, pretraining our model on image and miRNA data and finetuning on miRNA-only data can yield the best performance (e.g., the c-index of 67.29%), which is about 4% higher than the best baseline model on any combination of multimodal data. Even compared with our two-stage workflow PathOmics [2], which does not have contrastive learning in the pretraining scheme, our current method achieves better performance among the majority of data modality combinations. For example, under the same pretraining and finetuning data modality usage, we achieved about 9% performance improvement by pretraining the model with image and miRNA data and 1% performance improvement by pretraining on image and Methylation data. Such significant performance improvement demonstrates the efficiency of our contrastive-based pretraining strategy. Our contrastive-based pretraining is able to capture the complex interactions among multimodal data while distinguishing the difference between patients, while the pretraining scheme in PathOmics does not have the capability to learn patient differences.

5.4.3 Ablation Analysis

We verify the optimal design of model architecture and training parameters by designing two groups of ablation studies by comparing model performance with the various parameter settings: (1) the model performance of batch size and (2) the effect of the group of images. Different from the previous studies for survival outcome predictions (e.g., MCAT, PathOmics, etc.), which specify the batch size as 1, our proposed method is flexible to increase the batch size. Furthermore, as demonstrated in multiple contrastive-based methods, the larger the batch size is, the more available pairs will exist, and the conclusion could be different from our specific dataset. Unlike natural image-text datasets, the medical-domain dataset is not large enough. Hence, it could be crucial to select the proper batch size in our study to achieve the optimal performance of patient survival outcome prediction. Because of the total number of TCGA-COAD, we set the batch sizes as 2, 4, 8, and 16 for both the pertaining and finetuning stages. Our approach can yield the optimal performance by setting batch size as 8. We determined the optimal option for the number of batch sizes as 8 in our study, which yields the best performance among 2/3 types of multimodal combinations. Finally, we explore the effect of the group of images. We evaluated the different number of image groups, which vary in 2, 4, and 8. We found that the number of image groups can affect the model performance, as shown in Figure 5.3(b). We determined the optimal option for the number of image groups is 4 in our study, which yields the best performance among 2/3 types of multimodal combinations.

5.5 Summary

With the emergence of multimodal learning in the medical domain, it is promising to develop the domain-specific multimodal model to enhance the survival prognosis of colorectal cancer patients across various clinical data scenarios. Our study demonstrates the capability of the proposed contrastive-based PathOmics model in

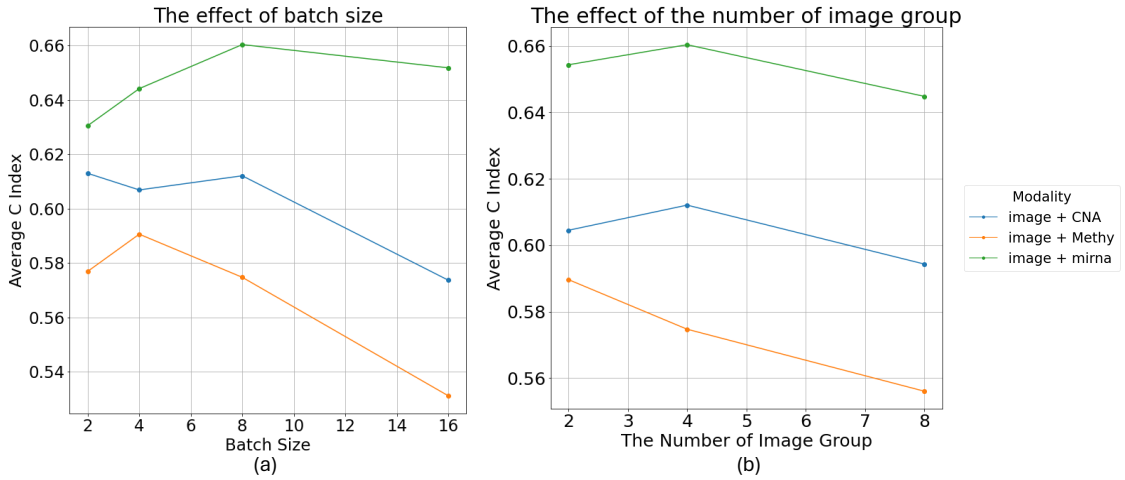


Figure 5.3: Ablation study. In (a) and (b), we evaluate the effect of using various batch size and the number of image group on TCGA-CRC. We show the average C-index among three times of running in the figure.

improving the survival prediction performance for patients with colorectal cancer. This approach opens the way for investigating the fundamental insights into the intricate genotype-phenotype interactions present in complex cancer data across different modalities. Furthermore, we improved the design of the previous study to enable a robust prediction of survival outcomes by introducing contrastive-based multimodal learning to distinguish patient differences. The modality-flexible finetuning strategy opens the possibility of broadening the usage of the dataset. In the real-world application scenario, the proposed modality-flexible finetuning strategy is able to introduce benefits to patients with missing data. In the stage of single modality finetuning, even when only utilizing image data, the model demonstrates the capability to generate genomic-related image embeddings. The reason is that the multimodal knowledge aggregation has been acquired during the model pretraining phase. Consequently, our cross-modal information aggregation alleviates the strict modality constraints during the finetuning stage.

CHAPTER 6: CONCLUSIONS AND FUTURE WORKS

Exploring histopathological and genomics data promises to enhance our understanding of complex cancer biology, enabling a better patient outcome assessment. In this dissertation, we proposed a graph neural network (GNN) framework that allows multi-region spatial connection of tiles to predict cross-scale molecular profile status in colon cancer. We demonstrated the validity of spatial connections of tumor tiles built upon the geometric coordinate from the raw whole-slide images (WSI) that were not reported in prior studies. We provided the interpretation by visualizing the image tiles and measuring the topological structure of tile-connected graphs. The explorations broadened our understanding of histopathological characteristics, establishing connections to a wide range of cross-scale molecular profile alterations, ranging from gene mutations and copy number alterations to functional protein expressions indicative of treatment relevance.

To better explore and utilize the inherent interaction among histopathological and genomics data, developing multimodal workflow becomes crucial to advance the survival assessment of cancer patients in real-world clinical applications. We demonstrated that the proposed PathOmics framework is useful for improving the survival prediction of colon and rectum cancer patients compared with the conventional attention-based and multimodal-based state-of-the-art models. Unsupervised pretraining reduces the dependency on data annotation and opens up perspectives for extracting and understanding the intrinsic genotype-phenotype interactions hidden in complex cancer data across modalities. Yet, the patient distinguishes the challenge that remains to be addressed. We demonstrated that the proposed contrastive-based PathOmics framework is useful for improving the survival prediction of colon and

rectum cancer patients. We further improved the design of the previous workflow to enable a robust prediction of survival outcomes by introducing a contrastive-based multimodal pretraining strategy to distinguish patient differences when aggregating the multimodal knowledge. Furthermore, the efficient finetuning approach broadens the scope of dataset usage, particularly for model finetuning and evaluation. The proposed modality-flexible finetuning strategy introduces the possibility to patients whose data has missing modalities. In single-modality finetuning, the single-modal data can still utilize the complementary knowledge from other modalities. For example, even if the patient in finetuning only has image-based data, the model is able to generate genomic-related image feature embedding for patient survival outcome prediction. The complex multimodal knowledge aggregation has already been finished in the model pretraining stage. Hence, cross-modal information aggregation reduces the requirement for data modality in the finetuning stage.

To fully leverage the inherent interactions between histopathological and genomics data, the development of multimodal workflow is important for advancing the survival assessment of cancer patients in real-world clinical scenarios. Our study demonstrates that the proposed PathOmics framework significantly enhances the survival prediction performance for colorectal cancer patients compared to conventional attention-based and multimodal-based state-of-the-art models. Leveraging an unsupervised pertaining scheme allows us to reduce reliance on data annotation and facilitate the intrinsic genotype-phenotype interactions present in complex cancer data across different modalities. However, the patient distinction remains a challenge that needs to be addressed in the PathOmics workflow. To overcome this challenge, our contrastive-based PathOmics framework introduces a robust strategy for survival outcome prediction by emphasizing patient differences during multimodal knowledge aggregation. Moreover, our proposed workflow continues using a modal-flexible finetuning approach, expanding the usage of datasets for model finetuning and evaluation. The modality-flexible

finetuning strategy accommodates patients with missing modalities, enabling a more inclusive approach to survival prediction. The proposed modality-flexible finetuning strategy introduces the possibility to patients whose data has missing modalities. In single-modality finetuning, the single-modal data can still utilize the complementary knowledge from other modalities. For instance, even if the patient in finetuning only has image-based data, the model is capable of generating genomic-related image feature embeddings for patient survival outcome prediction. The multimodal knowledge aggregation has already been done in the pretraining stage.

As a limitation of the proposed pathology-and-genomics for patient outcome prediction pipelines in this dissertation, we acknowledge that we only perform and validate our models on colorectal cancer in this dissertation. It is reasonable to evaluate the generalization power of our proposed pipelines by extending the scope across different types of cancers in the future. For example, we would like to extend our pretraining scheme on the multimodal pan-cancer dataset to aggregate multiple disease patterns, enabling a comprehensive understanding of human cancer. The potential interactions among cancers may also introduce novel benefits to patient outcome analysis. Finally, we plan to make more efforts on our methodology to enhance the performance of patient outcome prediction, including introducing recent powerful foundation models enabling a better capability in disease pattern capturing [97, 98, 99]. Also, a general-purpose foundation model can be a promising application in the healthcare domain. In this dissertation, all proposed model architectures focused on solving the unique question, e.g., molecular profile alteration and patient survival outcome prediction. Yet, in the real-world clinical workflow, the patient does not have to find two doctors to answer these questions, while a single doctor can provide the answers to multiple questions to the patient. In the future, a general-purpose foundation model in healthcare can help accelerate the real-world clinical workflow by producing the answer to multiple tasks.

REFERENCES

[1] K. Ding, M. Zhou, H. Wang, S. Zhang, and D. N. Metaxas, "Spatially aware graph neural networks and cross-level molecular profile prediction in colon cancer histopathology: a retrospective multi-cohort study," *The Lancet Digital Health*, vol. 4, no. 11, pp. e787–e795, 2022.

[2] K. Ding, M. Zhou, D. N. Metaxas, and S. Zhang, "Pathology-and-genomics multimodal transformer for survival outcome prediction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 622–631, Springer, 2023.

[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[4] M. Yang, H. Yang, L. Ji, X. Hu, G. Tian, B. Wang, and J. Yang, "A multi-omics machine learning framework in predicting the survival of colorectal cancer patients," *Computers in Biology and Medicine*, vol. 146, p. 105516, 2022.

[5] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, "Deep learning–based multi-omics integration robustly predicts survival in liver cancerusing deep learning to predict liver cancer prognosis," *Clinical Cancer Research*, vol. 24, no. 6, pp. 1248–1259, 2018.

[6] H. Ma, F. Jiang, Y. Rong, Y. Guo, and J. Huang, "Toward robust self-training paradigm for molecular prediction tasks," *Journal of Computational Biology*, vol. 31, no. 3, pp. 213–228, 2024.

[7] H. Ma, F. Jiang, Y. Rong, Y. Guo, and J. Huang, "Robust self-training strategy for various molecular biology prediction tasks," in *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–5, 2022.

[8] Y. Guo, J. Wu, H. Ma, and J. Huang, "Self-supervised pre-training for protein embeddings using tertiary structures," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, pp. 6801–6809, 2022.

[9] H. Ma, Y. Bian, Y. Rong, W. Huang, T. Xu, W. Xie, G. Ye, and J. Huang, "Cross-dependent graph neural networks for molecular property prediction," *Bioinformatics*, vol. 38, no. 7, pp. 2003–2009, 2022.

[10] A. J. Gentles, A. M. Newman, C. L. Liu, S. V. Bratman, W. Feng, D. Kim, V. S. Nair, Y. Xu, A. Khuong, C. D. Hoang, *et al.*, "The prognostic landscape of genes and infiltrating immune cells across human cancers," *Nature medicine*, vol. 21, no. 8, pp. 938–945, 2015.

[11] S. Ramón y Cajal, M. Sesé, C. Capdevila, T. Aasen, D. Mattos-Arruda, S. J. Diaz-Cano, J. Hernández-Losa, J. Castellví, *et al.*, "Clinical implications of intratumor heterogeneity: challenges and opportunities," *Journal of Molecular Medicine*, vol. 98, no. 2, pp. 161–177, 2020.

[12] T. Armaghany, J. D. Wilson, Q. Chu, and G. Mills, "Genetic alterations in colorectal cancer," *Gastrointestinal cancer research: GCR*, vol. 5, no. 1, p. 19, 2012.

[13] J. Li, Y. Lu, R. Akbani, Z. Ju, P. L. Roebuck, W. Liu, J.-Y. Yang, B. M. Broom, R. G. Verhaak, D. W. Kane, *et al.*, "Tcpa: a resource for cancer functional proteomics data," *Nature methods*, vol. 10, no. 11, pp. 1046–1047, 2013.

[14] R. Akbani, P. K. S. Ng, H. M. Werner, M. Shahmoradgoli, F. Zhang, Z. Ju, W. Liu, J.-Y. Yang, K. Yoshihara, J. Li, *et al.*, "A pan-cancer proteomic perspective on the cancer genome atlas," *Nature communications*, vol. 5, no. 1, pp. 1–15, 2014.

[15] M. Chen, B. Zhang, W. Topatana, J. Cao, H. Zhu, S. Juengpanich, Q. Mao, H. Yu, and X. Cai, "Classification and mutation prediction based on histopathology h&e images in liver cancer using deep learning," *NPJ precision oncology*, vol. 4, no. 1, pp. 1–7, 2020.

[16] J. De Matos, A. d. S. Britto Jr, L. E. Oliveira, and A. L. Koerich, "Histopathologic image processing: A review," *arXiv preprint arXiv:1904.07900*, 2019.

[17] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE reviews in biomedical engineering*, vol. 2, pp. 147–171, 2009.

[18] A. Marusyk, V. Almendro, and K. Polyak, "Intra-tumour heterogeneity: a looking glass for cancer?," *Nature reviews cancer*, vol. 12, no. 5, pp. 323–334, 2012.

[19] J. N. Kather, A. T. Pearson, N. Halama, D. Jäger, J. Krause, S. H. Loosen, A. Marx, P. Boor, F. Tacke, U. P. Neumann, *et al.*, "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer," *Nature medicine*, vol. 25, no. 7, pp. 1054–1056, 2019.

[20] M. Bilal, S. E. A. Raza, A. Azam, S. Graham, M. Ilyas, I. A. Cree, D. Snead, F. Minhas, and N. M. Rajpoot, "Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study," *The Lancet Digital Health*, vol. 3, no. 12, pp. e763–e772, 2021.

[21] H. Qu, M. Zhou, Z. Yan, H. Wang, V. K. Rustgi, S. Zhang, O. Gevaert, and D. N. Metaxas, "Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning," *NPJ precision oncology*, vol. 5, no. 1, pp. 1–11, 2021.

[22] K. Ding, Q. Liu, E. Lee, M. Zhou, A. Lu, and S. Zhang, "Feature-enhanced graph networks for genetic mutational prediction using histopathological images in colon cancer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 294–304, Springer, 2020.

[23] K. Ding, M. Zhou, Z. Wang, Q. Liu, C. W. Arnold, S. Zhang, and D. N. Metaxas, "Graph convolutional networks for multi-modality medical imaging: Methods, architectures, and clinical applications," *arXiv preprint arXiv:2202.08916*, 2022.

[24] A. Cheerla and O. Gevaert, "Deep learning with multimodal representation for pancancer prognosis prediction," *Bioinformatics*, vol. 35, no. 14, pp. i446–i454, 2019.

[25] R. J. Chen, M. Y. Lu, J. Wang, D. F. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood, "Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 4, pp. 757–770, 2020.

[26] R. J. Chen, M. Y. Lu, D. F. Williamson, T. Y. Chen, J. Lipkova, Z. Noor, M. Shaban, M. Shady, M. Williams, B. Joo, *et al.*, "Pan-cancer integrative histology-genomic analysis via multimodal deep learning," *Cancer Cell*, vol. 40, no. 8, pp. 865–878, 2022.

[27] R. J. Chen, M. Y. Lu, W.-H. Weng, T. Y. Chen, D. F. Williamson, T. Manz, M. Shady, and F. Mahmood, "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4025, 2021.

[28] X. Xing, Z. Chen, M. Zhu, Y. Hou, Z. Gao, and Y. Yuan, "Discrepancy and gradient-guided multi-modal knowledge distillation for pathological glioma grading," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pp. 636–646, Springer, 2022.

[29] J. N. Kather, L. R. Heij, H. I. Grabsch, C. Loeffler, A. Echle, H. S. Muti, J. Krause, J. M. Niehues, K. A. Sommer, P. Bankhead, *et al.*, "Pan-cancer image-based detection of clinically actionable genetic alterations," *Nature Cancer*, vol. 1, no. 8, pp. 789–799, 2020.

[30] B. Schmauch, A. Romagnoni, E. Pronier, C. Saillard, P. Maillé, J. Calderaro, A. Kamoun, M. Sefta, S. Toldo, M. Zaslavskiy, *et al.*, "A deep learning model to predict rna-seq expression of tumours from whole slide images," *Nature communications*, vol. 11, no. 1, pp. 1–15, 2020.

[31] A. Echle, H. I. Grabsch, P. Quirke, P. A. van den Brandt, N. P. West, G. G. Hutchins, L. R. Heij, X. Tan, S. D. Richman, J. Krause, *et al.*, "Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning," *Gastroenterology*, vol. 159, no. 4, pp. 1406–1416, 2020.

[32] R. Yamashita, J. Long, T. Longacre, L. Peng, G. Berry, B. Martin, J. Higgins, D. L. Rubin, and J. Shen, "Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study," *The Lancet Oncology*, vol. 22, no. 1, pp. 132–141, 2021.

[33] Y. Fu, A. W. Jung, R. V. Torne, S. Gonzalez, H. Vöhringer, A. Shmatko, L. R. Yates, M. Jimenez-Linan, L. Moore, and M. Gerstung, "Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis," *Nature Cancer*, vol. 1, no. 8, pp. 800–810, 2020.

[34] W. Lu, S. Graham, M. Bilal, N. Rajpoot, and F. Minhas, "Capturing cellular topology in multi-gigapixel pathology images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 260–261, 2020.

[35] D. Reisenbüchler, S. J. Wagner, M. Boxberg, and T. Peng, "Local attention graph-based transformer for multi-target genetic alteration prediction," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II*, pp. 377–386, Springer, 2022.

[36] J. Lipkova, R. J. Chen, B. Chen, M. Y. Lu, M. Barbieri, D. Shao, A. J. Vaidya, C. Chen, L. Zhuang, D. F. Williamson, *et al.*, "Artificial intelligence for multimodal data integration in oncology," *Cancer Cell*, vol. 40, no. 10, pp. 1095–1110, 2022.

[37] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[38] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[39] Y. Zhou, S. Graham, N. Alemi Koohbanani, M. Shaban, P.-A. Heng, and N. Rajpoot, "Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[40] T. D. Cook, D. T. Campbell, and W. Shadish, *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA, 2002.

[41] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," *arXiv preprint arXiv:1810.00826*, 2018.

[42] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[43] K. Xu, C. Li, Y. Tian, T. Sonobe, K.-i. Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," in *International Conference on Machine Learning*, pp. 5453–5462, PMLR, 2018.

[44] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[45] N. Zamanitajeddin, M. Jahanifar, and N. Rajpoot, "Cells are actors: Social network analysis with classical ml for sota histology image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 288–298, Springer, 2021.

[46] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (tcga): an immeasurable source of knowledge," *Contemporary oncology*, vol. 19, no. 1A, p. A68, 2015.

[47] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[48] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, *et al.*, "Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal," *Science signaling*, vol. 6, no. 269, pp. pl1–pl1, 2013.

[49] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *2009 IEEE international symposium on biomedical imaging: from nano to macro*, pp. 1107–1110, IEEE, 2009.

[50] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[51] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, *et al.*, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS medicine*, vol. 16, no. 1, p. e1002730, 2019.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[53] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, pp. 719–724, 2009.

[54] D. Soulières, W. Greer, A. M. Magliocco, D. Huntsman, S. Young, M.-S. Tsao, and S. Kamel-Reid, "Kras mutation testing in the treatment of metastatic colorectal cancer with anti-egfr therapies," *Current Oncology*, vol. 17, no. s1, pp. 31–40, 2010.

[55] D. S. Williams, D. Mouradov, C. Browne, M. Palmieri, M. J. Elliott, R. Nightingale, C. G. Fang, R. Li, J. M. Mariadason, I. Faragher, *et al.*, "Overexpression of tp53 protein is associated with the lack of adjuvant chemotherapy benefit in patients with stage iii colorectal cancer," *Modern Pathology*, vol. 33, no. 3, pp. 483–495, 2020.

[56] D. Stuart and W. R. Sellers, "Linking somatic genetic alterations in cancer to therapeutics," *Current opinion in cell biology*, vol. 21, no. 2, pp. 304–310, 2009.

[57] D. M. Oliveira, G. Santamaria, C. Laudanna, S. Migliozzi, P. Zoppoli, M. Quist, C. Grasso, C. Mignogna, L. Elia, M. C. Faniello, *et al.*, "Identification of copy number alterations in colon cancer from analysis of amplicon-based next generation sequencing data," *Oncotarget*, vol. 9, no. 29, p. 20409, 2018.

[58] M. Frattini, P. Saletti, E. Romagnani, V. Martin, F. Molinari, M. Ghisletta, A. Camponovo, L. Etienne, F. Cavalli, and L. Mazzucchelli, "Pten loss of expression predicts cetuximab efficacy in metastatic colorectal cancer patients," *British journal of cancer*, vol. 97, no. 8, pp. 1139–1145, 2007.

[59] M. Reschke, D. Mihic-Probst, E. H. Van Der Horst, P. Knyazev, P. J. Wild, M. Hutterer, S. Meyer, R. Dummer, H. Moch, and A. Ullrich, "Her3 is a determinant for poor prognosis in melanoma," *Clinical Cancer Research*, vol. 14, no. 16, pp. 5188–5197, 2008.

[60] A. Lagree, M. Mohebpour, N. Meti, K. Saednia, F.-I. Lu, E. Slodkowska, S. Gandhi, E. Rakovitch, A. Shenfield, A. Sadeghi-Naini, *et al.*, "A review and comparison of breast tumor cell nuclei segmentation performances using deep convolutional neural networks," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.

[61] T. Qing, S. Zhu, C. Suo, L. Zhang, Y. Zheng, and L. Shi, "Somatic mutations in zfhx4 gene are associated with poor overall survival of chinese esophageal squamous cell carcinoma patients," *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[62] W. Ge, H. Hu, W. Cai, J. Xu, W. Hu, X. Weng, X. Qin, Y. Huang, W. Han, Y. Hu, *et al.*, "High-risk stage iii colon cancer patients identified by a novel five-gene mutational signature are characterized by upregulation of il-23a and gut bacterial translocation of the tumor microenvironment," *International journal of cancer*, vol. 146, no. 7, pp. 2027–2035, 2020.

[63] J. Noorbakhsh, S. Farahmand, S. Namburi, D. Caruana, D. Rimm, M. Soltaniehha, K. Zarringhalam, J. H. Chuang, *et al.*, "Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images," *Nature communications*, vol. 11, no. 1, pp. 1–14, 2020.

[64] A. D. Powers and S. P. Palecek, "Protein analytical assays for diagnosing, monitoring, and choosing treatment for cancer patients," *Journal of healthcare engineering*, vol. 3, no. 4, pp. 503–534, 2012.

[65] D. Li, C. Lin, N. Li, Y. Du, C. Yang, Y. Bai, Z. Feng, C. Su, R. Wu, S. Song, *et al.*, "Plagl2 and pofut1 are regulated by an evolutionarily conserved bidirectional promoter and are collaboratively involved in colorectal cancer by maintaining stemness," *EBioMedicine*, vol. 45, pp. 124–138, 2019.

[66] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2020.

[67] S. Vickovic, G. Eraslan, F. Salmén, J. Klughammer, L. Stenbeck, D. Schapiro, T. Äijö, R. Bonneau, L. Bergenstråhle, J. F. Navarro, *et al.*, "High-definition spatial transcriptomics for in situ tissue profiling," *Nature methods*, vol. 16, no. 10, pp. 987–990, 2019.

[68] M. J. Żelazowski, E. Płuciennik, G. Pasz-Walczak, P. Potemski, R. Kordek, and A. K. Bednarek, "Wwox expression in colorectal cancer—a real-time quantitative rt-pcr study," *Tumor Biology*, vol. 32, no. 3, pp. 551–560, 2011.

[69] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, "The molecular signatures database hallmark gene set collection," *Cell systems*, vol. 1, no. 6, pp. 417–425, 2015.

[70] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*, pp. 2127–2136, PMLR, 2018.

[71] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[72] S. Kirk, Y. Lee, C. Sadow, S. Levine, C. Roche, E. Bonaccio, and J. Filiippini, "Radiology data from the cancer genome atlas colon adenocarcinoma [tcga-coad] collection," *The Cancer Imaging Archive*, 2016.

[73] S. Kirk, Y. Lee, C. Sadow, and Levine, "The cancer genome atlas rectum adenocarcinoma collection (tcga-read) (version 3) [data set]," *The Cancer Imaging Archive*, 2016.

[74] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, *et al.*, "The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data," *Cancer discovery*, vol. 2, no. 5, pp. 401–404, 2012.

[75] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[76] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," *Advances in neural information processing systems*, vol. 10, 1997.

[77] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," *Advances in neural information processing systems*, vol. 30, 2017.

[78] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in neural information processing systems*, vol. 34, pp. 2136–2147, 2021.

[79] K. Ding, M. Zhou, H. Wang, O. Gevaert, D. Metaxas, and S. Zhang, "A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer," *Scientific Data*, vol. 10, no. 1, p. 231, 2023.

[80] Y. Gao, Z. Li, D. Liu, M. Zhou, S. Zhang, and D. N. Meta, "Training like a medical resident: Universal medical image segmentation via context prior learning," *arXiv preprint arXiv:2306.02416*, 2023.

[81] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020.

[82] P. Hager, M. J. Menten, and D. Rueckert, "Best of both worlds: Multimodal contrastive learning with tabular and imaging data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23924–23935, 2023.

[83] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[84] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.

[85] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[86] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[87] Z. Zhao, Y. Liu, H. Wu, Y. Li, S. Wang, L. Teng, D. Liu, X. Li, Z. Cui, Q. Wang, *et al.*, "Clip in medical imaging: A comprehensive survey," *arXiv preprint arXiv:2312.07353*, 2023.

[88] Z. Wu, Y. Xiong, S. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *CVPR*, 2018.

[89] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[90] J. Yang, C. Li, P. Zhang, B. Xiao, C. Liu, L. Yuan, and J. Gao, "Unified contrastive learning in image-text-label space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19163–19173, 2022.

[91] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 18661–18673, Curran Associates, Inc., 2021.

[92] A. Taleb, M. Kirchler, R. Monti, and C. Lippert, "Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20908–20921, 2022.

[93] W. Ko, W. Jung, E. Jeon, and H.-I. Suk, "A deep generative–discriminative learning for multimodal representation in imaging genetics," *IEEE Transactions on Medical Imaging*, vol. 41, no. 9, pp. 2348–2359, 2022.

[94] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[95] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[96] J. Kenton and L. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACLHLT*, pp. 4171–4186, 2019.

[97] Y. Zhang, J. Gao, Z. Tan, L. Zhou, K. Ding, M. Zhou, S. Zhang, and D. Wang, "Data-centric foundation models in computational healthcare: A survey," *arXiv preprint arXiv:2401.02458*, 2024.

[98] X. Wang, X. Zhang, G. Wang, J. He, Z. Li, W. Zhu, Y. Guo, Q. Dou, X. Li, D. Wang, *et al.*, "Openmedlab: An open-source platform for multi-modality foundation models in medicine," *arXiv preprint arXiv:2402.18028*, 2024.

[99] S. Zhang and D. Metaxas, "On the challenges and perspectives of foundation models for medical image analysis," *Medical Image Analysis*, p. 102996, 2023.

APPENDIX A: SUPPLEMENTARY TABLES

Table A.1: Gene mutation prediction results (reproduced with permission from Elsevier [1]). For TCGA-COAD, we summarized AUC (with 95% CI) of our graph networks models. For TCGA-READ, we trained the model on colon cancer (TCGA-COAD) and directly evaluated the gene mutation data on rectum cancer.

	TCGA-COAD		TCGA-READ	
Gene	Mutation percentage	Slide AUC (95% CIs)	Mutation percentage	Slide AUC (95% CIs)
APC	73.40%	82.19 (78.00, 86.25)	77.60%	66.70 (57.46, 75.60)
TP53	58.65%	81.68 (77.94, 85.50)	73.60%	68.02 (59.63, 76.30)
TTN	50.64%	79.48 (75.21, 83.44)	35.29%	63.71 (54.83, 71.74)
RYR1	44.55%	84.86 (81.06, 88.47)	40.00%	77.44 (70.31, 84.33)
KRAS	42.95%	80.16 (75.83, 83.93)	38.40%	71.02 (63.16, 77.67)
PIK3CA	32.37%	79.85 (75.14, 84.18)	21.60%	67.93 (58.22, 76.91)
SYNE1	28.53%	81.94 (75.21, 83.44)	21.60%	70.46 (61.37, 79.61)
MUC16	27.24%	79.94 (74.87, 84.51)	15.20%	75.27 (67.12, 82.76)
FAT4	24.68%	83.40 (79.02, 87.65)	14.40%	62.20 (50.29, 73.49)
OBSCN	20.51%	82.52 (77.43, 87.22)	8.80%	77.09 (63.43, 89.47)
ZFHX4	19.87%	83.17 (78.00, 87.98)	7.20%	81.80 (72.20, 89.70)
RYR2	19.55%	87.08 (83.28, 90.82)	15.20%	66.14 (55.56, 76.45)
LRP1B	19.15%	84.18 (79.21, 88.39)	14.40%	66.70 (51.06, 71.62)
FBXW7	18.91%	82.95 (78.20, 87.03)	13.60%	65.01 (54.36, 75.79)
CSMD3	18.90%	82.17 (77.82, 86.57)	9.60%	70.58 (55.79, 84.83)
CSMD1	18.59%	84.18 (79.23, 88.97)	11.20%	66.22 (54.71, 76.96)
DNAH5	18.59%	83.75 (79.20, 87.65)	9.6%	66.00 (51.55, 79.46)
FLG	16.35%	82.54 (77.41, 87.14)	13.60%	70.91 (60.35, 82.70)
FAT3	15.38%	87.01 (82.03, 91.76)	10.40%	73.94 (63.29, 83.81)
DNAH11	15.06%	82.42 (77.16, 87.75)	9.60%	73.16 (59.23, 86.70)

Table A.2: Copy number alteration gene prediction results. For TCGA-COAD, we summarized AUC (with 95% CI) of our graph networks models (reproduced with permission from Elsevier [1]). For TCGA-READ, we trained the model on colon cancer (TCGA-COAD) and directly evaluated the gene CNA data on rectum cancer.

	TCGA-COAD		TCGA-READ	
Gene	Alteration percentage	Slide AUC (95% CIs)	Alteration percentage	Slide AUC (95% CIs)
CCSER1	23.08%	81.91 (77.16, 86.54)	17.60%	63.68 (51.51, 74.71)
COX4I2	14.42%	78.07 (71.51, 83.35)	6.40%	70.03 (45.67, 70.58)
CSMD1	11.22%	79.86 (73.08, 85.67)	8.00%	76.48 (64.78, 86.71)
DEFB118	9.29%	88.39 (83.51, 93.33)	11.20%	71.72 (57.49, 84.58)
DUSP15	8.10%	79.23 (71.28, 86.65)	15.20%	57.60 (44.60, 69.94)
FOXS1	8.65%	79.83 (73.18, 88.14)	8.80%	73.52 (62.13, 84.29)
ID1	8.01%	79.81 (69.27, 89.18)	15.20%	62.44 (50.55, 73.45)
MACROD2	8.01%	81.98 (73.34, 89.68)	15.20%	58.64 (47.29, 69.75)
MYLK2	8.01%	80.91 (72.65, 88.38)	15.20%	62.66 (48.24, 71.84)
HCK	7.69%	80.39 (72.31, 87.94)	15.20%	60.50 (48.24, 71.84)
KIF3B	7.69%	85.74 (78.36, 92.23)	15.20%	57.30 (44.86, 69.20)
PDRG1	7.69%	87.47 (80.16, 93.33)	14.40%	58.41 (45.67, 70.58)
PLAGL2	7.69%	90.55 (86.02, 94.89)	15.20%	59.36 (48.44, 70.33)
POFUT1	7.69%	87.99 (77.31, 92.24)	15.20%	57.45 (45.81, 68.64)
RBFOX1	7.69%	77.86 (78.69, 94.22)	14.40%	59.45 (46.86, 71.82)
REM1	7.37%	87.99 (80.25, 94.73)	14.40%	62.98 (50.64, 74.40)
TM9SF4	7.69%	86.79 (79.06, 93.54)	15.20%	58.02 (45.45, 70.94)
TPX2	7.69%	81.97 (74.04, 89.09)	14.40%	63.06 (50.49, 74.18)
TSPY26P	7.69%	86.89 (79.16, 94.28)	14.40%	59.01 (46.55, 70.42)
WWOX	7.69%	86.62 (78.96, 93.11)	14.40%	62.05 (48.94, 73.46)

Table A.3: Functional protein expression prediction results. For TCGA-COAD, we summarized AUC (with 95% CI) of our graph networks models (reproduced with permission from Elsevier [1]). For TCGA-READ, we trained the model on colon cancer (TCGA-COAD) and directly evaluated the protein expression prediction on rectum cancer.

	TCGA-COAD		TCGA-READ	
Protein	High level expression percentage	Slide AUC (95% CIs)	High level expression percentage	Slide AUC (95% CIs)
CMET _pY1235	52.91%	84.03 (79.69, 88.05)	52.94%	50.39 (41.37, 60.05)
P53	52.47%	86.41 (82.44, 90.19)	50.98%	49.65 (40.29, 58.89)
STAT3 _pY705	52.47%	85.57 (81.16, 89.44)	48.04%	53.89 (60.29, 87.45)
ACC1	50.67%	85.50 (81.33, 89.60)	50.00%	51.77 (42.53, 61.83)
BRCA2	50.67%	84.97 (80.97, 88.81)	50.98%	52.83 (42.96, 63.08)
BCL2	50.67%	83.54 (78.80, 87.51)	51.96%	59.79 (50.79, 68.57)
SRC_pY416	50.67%	86.23 (82.21, 90.20)	43.13%	53.89 (40.33, 59.26)
NOTCH1	50.22%	84.66 (80.34, 88.66)	50.00%	48.35 (38.97, 57.66)
PTEN	50.22%	86.01 (81.97, 90.06)	49.02%	50.65 (41.07, 59.94)
CMYC	50.22%	84.17 (79.72, 88.06)	53.92%	50.02 (40.69, 59.52)
ACC_pS79	50.22%	85.12 (80.97, 89.04)	50.98%	48.75 (39.30, 58.49)
EGFR	47.09%	84.00 (79.36, 88.36)	52.94%	53.26 (43.62, 63.01)
BRAF	49.78%	85.84 (81.68, 90.03)	50.00%	52.96 (43.69, 62.38)
ATM	49.78%	88.28 (84.43, 91.67)	52.94%	53.38 (43.87, 62.92)
ERALPHA	49.77%	85.80 (81.51, 89.77)	51.96%	57.16 (47.54, 66.46)
ARID1A	48.88%	85.77 (81.41, 89.57)	51.96%	50.90 (41.17, 61.50)
HER2	48.43%	84.46 (79.79, 88.76)	50.00%	51.92 (42.34, 61.19)
HER3	48.43%	885.59 (81.39, 89.48)	50.00%	39.81 (30.15, 49.15)
AMPKAL PHA_pY172	46.64%	81.94 (77.70, 86.55)	47.06%	56.06 (46.83, 65.28)
EGFR _pY1173	45.29%	89.64 (86.29, 93.19)	50.98%	48.65 (38.85, 58.17)

Table A.4: External validation on CPTAC-COAD for gene mutation prediction (reproduced with permission from Elsevier [1]). We trained the model on colon cancer (TCGA-COAD) and directly validated the gene mutation data on the same cancer cohort in CPTAC-COAD.

Gene	Mutation percentage	Slide AUC (95% CIs)
APC	73.00%	36.12 (25.79, 46.55)
TP53	51.00%	40.04 (30.34, 49.76)
TTN	49.00%	60.01 (50.64, 68.93)
KRAS	31.00%	44.06 (35.52, 53.46)
RYR1	18.00%	81.38 (79.63, 94.48)
PIK3CA	16.00%	72.47 (55.76, 85,84))
SYNE1	22.00%	68.85 (59.30, 78.23)
MUC16	33.00%	69.04 (59.36, 77.76)
FAT4	22.00%	53.35 (43.84, 62.97)
OBSCN	24.00%	79.39 (70.05, 87.91)
ZFHX4	16.00%	51.67 (40.72, 63.20)
RYR2	16.00%	50.26 (36.36, 64.45)
LRP1B	20.00%	69.59 (60.72, 78.10)
FBXW7	14.00%	49.50 (37.62, 60.92)
CSMD3	22.00%	82.90 (73.69, 90.71)
CSMD1	22.00%	38.64 (28.18, 49.49)
DNAH5	14.00%	76.16 (67.11, 83.55)
FLG	15.00%	73.45 (63.26, 83.25)
FAT3	21.00%	63.74 (52.92, 75.37)
DNAH11	12.00%	82.01 (74.16, 88.82)

Table A.5: External validation on CPTAC-COAD fo gene CNA prediction (reproduced with permission from Elsevier [1]). We trained the model on colon cancer (TCGA-COAD) and directly validated the gene CNA status on the same cancer cohort in CPTAC-COAD.

Gene	Alteration percentage	Slide AUC (95% CIs)
CCSER1	22.00%	78.50 (67.87, 87.34)
FOXS1	12.00%	86.08 (79.67, 91.74)
DEFB118	9.00%	62.39 (51.37, 73.76)
COX4I2	8.00%	43.75 (29.05, 57.29)
CSMD1	5.00%	51.16 (39.79, 63.27)