Deep learning methods for the integration of multi-omics and histopathology data for precision medicine in oncology

Hakim Benkirane

▶ **To cite this version:**

Hakim Benkirane. Deep learning methods for the integration of multi-omics and histopathology data for precision medicine in oncology. Cancer. Université Paris-Saclay, 2024. English. NNT : 2024UPASR022 . tel-04871130

HAL Id: tel-04871130
https://theses.hal.science/tel-04871130v1

Submitted on 7 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

université PARIS-SACLAY

Deep Learning Methods for the integration of multi-omics and histopathology data for precision medicine in oncology

Méthodes d'apprentissage profond pour l'intégration de données multi-omiques et de pathologie numérique pour la médecine de précision en oncologie

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 570, Santé Publique, EDSP
Spécialité de doctorat: Biostatistiques et Data Sciences
Graduate School : Santé Publique. Référent : Faculté de médecine

Thèse préparée dans les unités de recherche **CESP** (Université Paris-Saclay, Gustave Roussy, INSERM) et **MICS EA 4037 Mathématiques et Informatique pour la Complexité et les Systèmes** (Université Paris-Saclay, CentraleSupélec), sous la direction de **Stefan Michiels**, Docteur et la co-direction de **Paul-Henry Cournède**, Professeur

Thèse soutenue à Paris-Saclay, le 04 Décembre 2024, par

Hakim BENKIRANE

Composition du jury

Membres du jury avec voix délibérative

Arthur Tenenhaus Professeur, Université Paris-Saclay	Président
Manuela ZUCKNICK Professeur, Université d'Oslo	Rapporteur & Examinatrice
Xavier ALAMEDA-PINEDA HDR, Université Grenoble-Alpes	Rapporteur & Examinateur
Mireia Crispín Ortuzar Dr., Université de Cambridge	Examinatrice
Hervé Delingette Professeur, Inria, Epione Team	Examinateur
Laura Cantini Dr., Institut Pasteur	Examinatrice

THESE DE DOCTORAT
NNT : 2024UPASR022

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Titre: Méthodes d'apprentissage profond pour l'intégration de données multi-omiques et de pathologie numérique pour la médecine de précision en oncologie
Mots clés: Apprentissage Statistique, Multi-omique, Histopathologie, Multimodalité, Interprétabilité, Médecine de Précision

Résumé: La médecine de précision est une approche émergente pour le traitement et la prévention des maladies qui prend en compte la variabilité individuelle dans les gènes, l'environnement et le mode de vie. L'objectif est de prédire plus précisément quelles stratégies de traitement et de prévention pour une maladie particulière fonctionneront dans quels groupes de personnes. En oncologie, la médecine de précision s'accompagne d'une augmentation drastique des données collectées pour chaque individu, caractérisée par une grande diversité de sources de données. Par exemple, les patients recevant un traitement contre le cancer sont souvent soumis à un profilage moléculaire complet, en plus du profilage clinique et des images de pathologie anatomique. Par conséquent, l'intégration de données multimodales (images, cliniques, moléculaires) est une question critique pour permettre la définition de modèles prédictifs individuels. Cette thèse aborde le développement de modèles computationnels et de stratégies d'apprentissage capables de déchiffrer des interactions complexes et de haute dimension. Un accent significatif est également mis sur l'explicabilité de ces modèles pilotés par l'IA, assurant que les prédictions soient compréhensibles et cliniquement exploitables.

Title: Deep Learning Methods for the integration of multi-omics and histopathology data for precision medicine in oncology
Keywords: Statistical Learning, Multi-omics, Histopathology, Multimodality, Interpretability, Precision medicine

Abstract: Precision medicine is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle. The objective it to predict more accurately which treatment and prevention strategies for a particular disease will work in which groups of people. In oncology, precision medicine comes with a drastic increase in the data that is collected for each individual, characterized by a large diversity of data sources. Advanced cancer patients receiving cancer treatment, for instance, are often subject to a complete molecular profiling, on top of clinical profiling and pathology images. As a consequence, integration methods for multi-modal data (image, clinical, molecular) is a critical issue to allow the definition of individual predictive models. This thesis tackles the development of computational models and learning strategies adept at deciphering complex, high-dimensional interactions. A significant focus is also placed on the explainability of these AI-driven models, ensuring that predictions are understandable and clinically actionable.

In loving memory of Ahmed Benkirane

With this thesis, your legacy will go on...

Acknowledgments

Earning a PhD is not merely the outcome of hard work; it is the result of invaluable guidance, collaboration, and constant support from so many individuals.

This journey began with the indispensable guidance of my supervisor, Paul-Henry, who first welcomed me into the lab during my master's thesis. From that moment on, you became a pivotal figure in my academic life, and I cannot overstate how much your mentorship has shaped the course of these past three years. Your unwavering enthusiasm for my growth and the projects I pursued never ceased to motivate me, while the countless opportunities you offered to broaden my research horizons allowed me to evolve as a scholar. Your confidence in my abilities was not only empowering but also a crucial factor in my development, and for that, I am deeply thankful.

I am equally indebted to Stefan, whose remarkable generosity not only opened the door to this PhD program but also provided me with the essential knowledge and tools to thrive. Stefan, your influence on my research has been profound, and your constant support, both academically and personally, has been a bedrock of my success. I cannot thank you enough for the key role you played in helping me achieve my goals.

I want to extend my heartfelt thanks to my colleagues at the MICS lab, who stood by my side throughout these transformative three years. Your camaraderie and support made this journey all the more rewarding. A special mention goes to Quentin, my indispensable partner in the thesis project. Working closely with you, sharing ideas, and collaborating on every aspect of our research has been an invaluable experience. Your insights and dedication played a significant role in shaping the outcomes of our work.

The lab would not have been complete without the vibrant spirit of the Moroccan crew. Othmane, your wisdom and calm demeanor were always a source of guidance; Aaron, your sense of adventure brought a spark of excitement into our day-to-day routines; and Imane, your captivating stories filled the lab with laughter and warmth. Each of you contributed something unique that made our time together in the lab not just productive, but genuinely enjoyable. Thank you for being an essential part of this journey.

I am deeply grateful to my coworkers at Oncostat and Gustave Roussy for welcoming me with open arms and making my time with you truly memorable. From friendly afterwork competitions that brought out our competitive spirits to team outings that strengthened our bond, each moment added a layer of joy to the challenges of the PhD journey. Your warmth and camaraderie created an environment where work felt like a shared adventure, and for that, I am incredibly thankful.

This thesis would not have been possible without the invaluable contributions of brilliant researchers who have significantly shaped my academic path. I am especially thankful to Maria, Stergios, and Elsa, with whom I had the great privilege of collaborating on numerous projects. Your expertise, insights, and willingness to engage in thought-provoking discussions not only enriched my research but also expanded my understanding of the field in profound ways. Your influence will always remain an integral part of my academic and professional development.

On a more personal level, I am profoundly grateful to my roommate Louis. Sharing this journey with you has been a true blessing. Your presence brought a sense of stability and comfort during the

most challenging times, and knowing that we were facing similar hurdles made the experience far less isolating. From late-night discussions to navigating the pressures of research, your companionship was a pillar of support throughout these years.

To Nicolas, I owe a special thank you for enduring my endless complaints over the past three years. Your patience and understanding were invaluable, and you always listened without judgment, even when I was at my most frustrated. I also want to extend my heartfelt thanks to Adrien, Julien, Fabien, and Leandre. Your faith in my abilities, your constant encouragement, and your belief in my potential helped push me forward, especially during moments of doubt. Knowing that I had your support meant more to me than I can express, and I'm incredibly fortunate to have friends like you by my side.

I cannot overlook the invaluable presence of my friends Omar, Saad, Oussama, Walid, and especially Yasser. Over the past three years, Yasser, our friendship has evolved into something truly special. Your unwavering support, wisdom, and companionship have been a constant source of strength for me, and I deeply appreciate how we've helped each other navigate this journey. I'm certain that as we move forward, our bond will only continue to deepen, and we will keep pushing each other toward the realization of our dreams.

Lastly, my family has been the unwavering foundation of my entire journey, and I owe so much of who I am today to their love, guidance, and sacrifice. They have been there at every step, offering not only emotional support but also the wisdom and strength I needed to persevere, even in the most challenging moments. This thesis is not just the culmination of my academic efforts, but a reflection of their enduring belief in me and the opportunities they have helped create. Every accomplishment I have achieved is as much theirs as it is mine, and I will always be grateful for the countless ways they have shaped my life, both personally and professionally. Without their constant encouragement and understanding, this journey would not have been possible.

List of Figures

List of Tables

Abbreviations

AUC	Area Under ROC-Curve
VAE	Variational Autoencoder
AE	Autoencoder
C-index	Concordance Index
IBS	Integrated Brier Score
BS	Brier Score
CNV	Copy Number Variation
LFS	Leukemia-Free Survival
OS	Overall Survival
IPCW	Inverse Probability of Censoring Weighting
KM	Kaplan-Meier
KL	Kullback-Leibler
LASSO	Area Under ROC-Curve
VAE	Least Absolute Shrinkage and Selection Operator
TCGA	The Cancer Genome Atlas
IALT	The International Adjuvant Lung Cancer Trial
GDC	Genomic Data Commons
MMD	Maximum Mean Discrepency
NSCLC	Non-Small Cell Lung Cancer
MDS	Myelodysplatic Syndromes
BRCA	Breast Cancer Carcinoma
BLCA	Bladder Cancer Carcinoma
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
COAD	Colon Adenocarcinoma
GBM	Glioblastoma Multiforme
PRAD	Prostate Adenocarcinoma
KIRC	Kidney Renal Clear Cell Carcinoma
RNAseq	RNA sequencing

Contents

Résumé

Le cancer reste l'une des maladies les plus complexes et mortelles auxquelles la médecine moderne est confrontée. En raison de sa nature hétérogène, chaque type de cancer est unique non seulement entre les individus, mais aussi au sein même des tumeurs d'un même patient. Les multiples mutations génétiques, les altérations moléculaires et les interactions entre les cellules cancéreuses et leur microenvironnement rendent difficile une approche uniforme pour tous les patients. Ainsi, l'émergence de la médecine de précision constitue une avancée significative dans le traitement du cancer. Plutôt que d'administrer des thérapies standards, cette approche vise à personnaliser les soins en fonction des caractéristiques spécifiques de chaque patient, en s'appuyant sur l'analyse de données complexes et hétérogènes issues de sources multiples, telles que les données multi-omiques (génomiques, transcriptomiques, protéomiques, etc.) et les images histopathologiques sous forme de Whole Slide Images (WSIs). Cependant, l'un des principaux défis de la médecine de précision reste l'intégration efficace et l'analyse approfondie de ces données hétérogènes pour extraire des informations cliniquement exploitables.

Cette thèse s'inscrit dans ce cadre en proposant de nouvelles méthodes basées sur l'apprentissage profond pour l'intégration des données multi-omiques et des WSIs, tout en garantissant une explicabilité essentielle à leur adoption dans la pratique clinique. L'objectif principal de ce travail est de développer des modèles capables d'améliorer non seulement les prédictions concernant les sous-types tumoraux et la survie des patients, mais aussi de rendre ces modèles interprétables pour les cliniciens, assurant ainsi une véritable utilisation en clinique dans le cadre de la médecine personnalisée.

Le premier défi auquel cette thèse répond est la gestion de l'intégration des données multimodales. Les données omiques, issues de technologies avancées telles que la génomique, la transcriptomique ou la protéomique, offrent une vision détaillée du profil moléculaire des tumeurs, mais manquent souvent de la dimension spatiale et morphologique que peuvent fournir les images histopathologiques. Inversement, les WSIs capturent des informations sur la structure et l'organisation des tissus tumoraux, mais n'apportent pas d'informations moléculaires. L'un des grands défis est donc de combiner ces deux types de données, en intégrant à la fois les caractéristiques moléculaires fines et les

informations morphologiques complexes des tumeurs, de manière à offrir une vision plus globale et plus complète du cancer.

Un autre défi important est la complexité et la grande dimensionnalité de ces données. Les données multi-omiques peuvent contenir des milliers de variables, chacune représentant un aspect différent de l'état biologique de la tumeur. De plus, les WSIs sont des images de très haute résolution, contenant des millions de pixels et des détails spatiaux complexes. Le traitement simultané de ces données massives et hétérogènes pose donc des défis considérables en matière de calcul, d'efficacité et d'interprétation biologique.

L'une des contributions majeures de cette thèse est le développement de **CustOmics**, un cadre d'apprentissage profond spécifiquement conçu pour intégrer et analyser les données multi-omiques. CustOmics répond au défi de la grande dimensionnalité des données en utilisant des autoencodeurs variationnels (VAE) pour capturer des représentations latentes compactes, mais informatives, des données multi-omiques. Cela permet de conserver les informations biologiquement pertinentes tout en réduisant la complexité des données. En mettant l'accent sur la réduction dimensionnelle, CustOmics est capable d'améliorer la classification des sous-types tumoraux ainsi que la prédiction de la survie des patients en se basant sur des ensembles de données omiques massives. Appliqué aux données du projet *The Cancer Genome Atlas* (TCGA), CustOmics a montré des résultats prometteurs, notamment dans le cas du cancer du poumon, en identifiant des sous-types tumoraux spécifiques et en améliorant la précision des prédictions de survie par rapport aux méthodes d'intégration traditionnelles. En plus de ses performances prédictives, CustOmics offre une flexibilité qui lui permet d'intégrer d'autres types de données, telles que des données cliniques ou des informations sur l'environnement tumoral, permettant ainsi une analyse plus globale du cancer.

Les avantages de CustOmics sont doubles : d'une part, il permet de tirer parti de la richesse des données multi-omiques en identifiant des biomarqueurs pertinents et des caractéristiques moléculaires spécifiques qui peuvent guider la personnalisation des traitements. D'autre part, il réduit le risque de sur-apprentissage et de confusion souvent associés à l'utilisation de données massives et de grande dimensionnalité, tout en offrant des résultats interprétables et fiables.

L'analyse des images histopathologiques joue un rôle fondamental dans le diagnostic et la classification des cancers. Cependant, les Whole Slide Images (WSIs), en raison de leur taille et de leur complexité, sont difficiles à analyser de manière automatisée. Pour résoudre ce problème, cette thèse

propose **Hyper-AdaC**, un modèle basé sur une représentation hypergraphique des WSIs. Hyper-AdaC permet de segmenter les WSIs en sous-structures significatives, ce qui facilite l'analyse et la compréhension des relations morphologiques complexes au sein des tissus tumoraux. L'une des forces de cette approche est qu'elle capture les relations spatiales fines au sein des tumeurs, permettant ainsi de mieux comprendre comment l'organisation des cellules et des tissus peut influencer la progression du cancer et la réponse aux traitements. Hyper-AdaC a montré des résultats impressionnants, notamment dans la classification des sous-types de cancer et dans la prédiction de la survie des patients, en exploitant des informations qui échappent souvent aux méthodes d'analyse traditionnelles.

L'un des aspects les plus importants de l'adoption des modèles d'apprentissage profond dans la pratique clinique est leur explicabilité. Les médecins doivent être en mesure de comprendre comment et pourquoi un modèle a pris une décision spécifique pour pouvoir s'y fier en toute confiance. Pour répondre à cette nécessité, nous avons développé deux outils d'explicabilité : **H&Explainer** et **GMM-CeFlow**.

H&Explainer permet d'offrir des explications interprétables des prédictions réalisées par des modèles d'apprentissage profond appliqués aux WSIs et aux données omiques. Grâce à des techniques telles que les valeurs SHAP (Shapley Additive exPlanations), H&Explainer identifie les régions des images ou les caractéristiques moléculaires qui influencent le plus les prédictions, fournissant ainsi aux cliniciens une compréhension claire des décisions prises par le modèle. Cela est particulièrement utile pour des cas comme le cancer du sein ou du poumon, où H&Explainer a permis de mettre en évidence des régions de la tumeur initialement jugées non significatives, mais qui se sont révélées critiques dans la classification tumorale.

GMM-CeFlow, quant à lui, introduit une approche d'analyse contre-factuelle pour explorer les scénarios "et si". Il permet de modifier certaines caractéristiques des données (WSIs ou données omiques) et d'observer comment ces modifications affecteraient les prédictions du modèle. Ce cadre offre aux cliniciens la possibilité de comprendre les limites et les sensibilités des modèles, tout en testant des scénarios alternatifs pour optimiser les décisions thérapeutiques. Par exemple, il est possible de simuler l'ajout ou la suppression de mutations spécifiques pour évaluer leur impact sur le pronostic ou les résultats cliniques.

Une autre contribution importante de cette thèse est **Multimodal CustOmics**, une extension

de CustOmics qui permet l'intégration simultanée des données omiques et des WSIs dans un cadre unique. Cette approche multimodale combine les réseaux de neurones convolutifs (CNN) pour l'analyse des images histopathologiques et les autoencodeurs variationnels pour les données omiques. Cela permet de capturer à la fois les caractéristiques microscopiques (via les WSIs) et moléculaires (via les données omiques), offrant ainsi une vue d'ensemble plus complète du cancer. Les expérimentations montrent que cette approche améliore de manière significative la classification des sous-types tumoraux et la prédiction de la survie des patients, en particulier dans les cas de cancer du poumon.

Malgré les avancées réalisées dans cette thèse, certains défis demeurent. L'un des principaux défis est la généralisation des modèles à d'autres types de cancers et à des ensembles de données plus diversifiés. L'intégration de nouvelles modalités de données, telles que la transcriptomique spatiale ou les données de biopsie liquide, pourrait également contribuer à améliorer encore davantage la précision des modèles.

Enfin, l'un des grands défis futurs sera de traduire ces avancées méthodologiques en outils directement utilisables en pratique clinique. Cela nécessitera des collaborations étroites entre informaticiens, cliniciens et biologistes afin de valider ces modèles dans des environnements cliniques réels.

En conclusion, cette thèse apporte des contributions dans le domaine de l'intégration des données multi-omiques et des WSIs, ouvrant ainsi la voie à une médecine de précision plus performante et personnalisée. Les outils développés, tels que **CustOmics**, **Hyper-AdaC**, **H&Explainer**, **GMM-CeFlow**, et **Multimodal CustOmics**, offrent des solutions

Preface

This manuscript presents the findings of my PhD research, conducted in collaboration with the Oncostat and MICS teams. The focus is on exploring how methodological advances and clinical applications intersect to develop computational models for precision medicine, with a particular emphasis on multimodality. The work aims to bridge the fields of data science and applied oncology, offering insights for professionals in both areas by highlighting the role of multimodal approaches, their impact, challenges, and potential benefits, helping data scientists and clinicians better navigate this evolving field.

Outline

The first chapter of this thesis lays the groundwork by providing a detailed overview of how artificial intelligence (AI) can be applied to understand cancer biology. It explores fundamental principles in both fields, showing how they interact and complement each other. The goal of this chapter is to equip professionals from various disciplines with the knowledge needed to grasp how AI can be of use to understand the complexities of cancer biology. By introducing key concepts, it prepares the reader for a deeper exploration of AI's role and application in cancer biology throughout the thesis.

In the second chapter, we explore the process of creating representations for complex data sets and their application in integrating multiple modalities. This chapter concludes a review of the literature and classification, analyzing different approaches to representation learning. The objective is to explain the methods used to derive meaningful insights from intricate data structures. Additionally, we examine multimodal integration, evaluating various techniques for combining data from different sources. By assessing the strengths and weaknesses of each approach, we aim to gather insights that will inform future developments.

The third chapter addresses the challenge of merging multi-omics data, which is crucial for understanding and tackling this complex task. We begin by outlining the different types of data involved, then discuss the challenges of integrating multiple omics data. This chapter introduces new approaches for effectively combining various molecular data sets, offering strategies to overcome challenges related to data diversity and size. The chapter's main focus is on CustOmics, an approach

introduced in a 2023 paper in PLoS Computational Biology. We analyze CustOmics, highlighting its ability to combine multiple omics datasets. To illustrate its practical application, we present a clinical case study on Myelodysplastic Syndrome (MDS), conducted in collaboration with Dr. Elsa Bernard and colleagues from the Karolinska Institute and the MDS Consortium.

The fourth chapter shifts focus to histopathology imaging, examining its complexities, challenges, and promising developments. It begins with an analysis of how histopathology slides are represented, laying the foundation for further discussion. This chapter forms the basis for a forthcoming article for the ML4H symposium, where we aim to clarify the core principles of histopathology slide visualization, which is crucial for applying machine learning techniques in this area. Additionally, we explore explainability in histopathology imaging, comparing different approaches, including attention mechanisms and human-understandable features. By evaluating these methods, we discuss their effectiveness in terms of explainability, outlining their respective strengths and limitations.

The fifth chapter concludes this investigation by focusing on the integration of multi-omics and histopathology data to enhance predictive models and improve our understanding of the biological mechanisms underlying cancer. The chapter expands on the methodology from the original paper, developing it into a broader framework called Multimodal CustOmics. This approach uses biological data to create robust and interpretable representations, ultimately advancing our understanding of disease mechanisms and supporting personalized treatment strategies. The primary aim of this study is to explore the levels of interpretability offered by the Multimodal CustOmics framework. By examining the methodology, we highlight how it can provide valuable insights into the biological mechanisms driving disease progression. The chapter centers on the application of Multimodal CustOmics in the International Adjuvant Lung Cancer Trial (IALT). By using real clinical data, we demonstrate the method's practical and clinical significance, showcasing its potential to improve prognostic and diagnostic outcomes.

Leveraging the progress in multimodal integration, I explored new spatial technologies to broaden the range of imaging and molecular interaction analysis. This effort resulted in a collaboration with Quentin Blampey to develop Novae, a model focused on spatial domain assignment and analysis, where I contributed by benchmarking state-of-the-art models and revealed issues in terms of batch-effect correction and clustering when dealing with multiple gene panels. Novae will serve as a key foundation for expanding its application to histopathology slide analysis and will eventually be inte-

grated into CustOmics to further advance multimodal integration capabilities.

Contributions

Published Papers

- **CustOmics: A versatile deep-learning based strategy for multi-omics integration**:
  - *PLoS Computational Biology*, 2023, 19(3), e1010921
  - **Authors**: Hakim Benkirane, Yoann Pradat, Stefan Michiels, Paul-Henry Cournède
  - **DOI**: https://doi.org/10.1371/journal.pcbi.1010921
- **Hyper-AdaC: adaptive clustering-based hypergraph representation of whole slide images for survival analysis**:
  - *Machine Learning for Health*, 2022, PMLR, 405-418.
  - **Authors**: Hakim Benkirane, Maria Vakalopoulou, Stergios Christodoulidis, Ingrid-Judith Garberis, Stefan Michiels, Paul-Henry Cournède
  - **DOI**: https://proceedings.mlr.press/v193/benkirane22a.html
- **Multimodal CustOmics: A Unified and Interpretable Multi-Task Deep Learning Framework for Multimodal Integrative Data Analysis in Oncology**:
  - *Biorxiv*, 2024, Cold Spring Harbor Laboratory.
  - **Authors**: Hakim Benkirane, Maria Vakalopoulou, David Planchard, Julien Adam, Ken Olaussen, Stefan Michiels, Paul-Henry Cournède
  - **DOI**: https://doi.org/10.1101/2024.01.20.576363
- **Novae: a graph-based foundation model for spatial transcriptomics data**:
  - *Biorxiv*, 2024, Cold Spring Harbor Laboratory.
  - **Authors**: Quentin Blampey, Hakim Benkirane, Nadège Bercovici, Fabrice André, Paul-Henry Cournède
  - **DOI**: https://doi.org/10.1101/2024.09.09.612009

Communications

- **Réseaux de neurones et intégration multi-omique pour la survie: quelles stratégies pour un meilleur apprentissage de représentation?**:
  - *EPICLIN*, 2022, 70, Revue d'Épidémiologie et de Santé Publique.
  - **Type**: Poster Presentation
  - **Authors**: Hakim Benkirane, Yoann Pradat, Stefan Michiels, Paul-Henry Cournède
- **Counterfactual Analysis for Digital Histopathology Slides Using Human Interpretable Features**:
  - *MIDL*, 2024, Paris.
  - **Type**: Poster Presentation
  - **Authors**: Hakim Benkirane, Maria Vakalopoulou, Stefan Michiels, Paul-Henry Cournède, William Lotter
- **Multimodal CustOmics: A Unified and Interpretable Multi-Task Deep Learning Framework for Multimodal Integrative Data Analysis in Oncology**:
  - *TIA Seminar*, 2024, Warwick University.
  - **Type**: Oral Presentation
  - **Authors**: Hakim Benkirane, Maria Vakalopoulou, David Planchard, Julien Adam, Ken Olaussen, Stefan Michiels, Paul-Henry Cournède

Introduction

Contents

Abstract

Cancer is a complex disease influenced by genetic and environmental factors. By leveraging advanced technologies to analyze traditional tissue samples and diverse biological data, we can gain new insights into the molecular processes driving cancer formation, ultimately leading to the development of personalized treatment strategies. Statistical learning is crucial for deciphering complex datasets, enabling the identification of patterns, and making predictions. However, ensuring that these models are explainable is essential to gain acceptance in clinical settings. This overview delves into how deep learning can tackle these challenges by enhancing the interpretability of models, thus bridging the gap between data integration and practical clinical application. We focus on presenting a framework that enhances cancer diagnosis and treatment while meeting the demands for clinical relevance and transparency.

1.1 . Introduction to Cancer: A multi-faceted disease

Cancer remains one of the most complex and challenging illnesses for modern medicine to address due to its wide variety of causes, symptoms, and effects on human health [113]. Essentially, cancer involves the uncontrolled growth of cells, leading to the formation of tumors that can become cancerous, invade nearby tissues, and sometimes spread to other parts of the body [77]. Benign tumors are less harmful as they do not invade surrounding tissues. In contrast, malignant tumors pose a serious health risk, potentially causing body dysfunction and leading to fatal consequences [**?**]. The complexity of understanding cancer results from its occurrence in various tissues and organs in mammals and the diverse symptoms and progression it presents. However, collaboration among clinicians, biologists, statisticians, and researchers from different fields has led to significant progress in our understanding of cancer, aiding in categorizing the disease, clarifying its processes, and improving treatment options. Decades of research and clinical practice have substantially transformed our approach to cancer care, significantly improving numerous patients' outlook, life expectancy, and well-being. By continually researching and developing new treatments, there is a growing sense of hope for achieving long-term remission and even curing more types of cancer, signaling a positive direction in the fight against this challenging illness.

1.1.1 . About Cancer

Cancer affects communities in every country to varying extents, as reported by cancer registries around the world [224]. Sadly, each year, millions of new cases of cancer are diagnosed, resulting in a substantial number of deaths worldwide. The occurrence of this disease is not uniform; it varies significantly due to factors such as location, environment, lifestyle, genetic predispositions, and economic status [145]. The most prevalent types of cancer in a specific area may indicate a complex interplay of these factors, which can provide valuable insights for public health strategies and research priorities.

The critical aspect of cancer's biological process is the occurrence of oncogenesis or tumorigenesis, which is the transformation of healthy cells into cancerous ones [146]. This transformation takes place through a series of steps, often referred to as a progression through different stages, beginning with initiation, then promotion, and ending with progression. The initial phase involves DNA changes in normal cells due to environmental factors, inherited mutations, or errors in DNA replication. The next step is promotion, during which these initial cells multiply excessively due to additional genetic alterations or in response to hormonal or growth factors. Finally, in the progression phase, cancer cells develop more aggressive characteristics, such as the ability to invade nearby tissues and create metastases.

Cancer disrupts normal cellular functions in various ways. One of its primary methods is bypassing the body's growth control mechanisms, leading to uncontrolled cell division. Additionally, cancer cells are capable of evading apoptosis, allowing them to survive beyond their expected lifespan [80]. They also stimulate angiogenesis, forming new blood vessels to nourish the growing tumor with necessary nutrients and oxygen. Furthermore, cancer cells employ immune evasion strategies, preventing detection and destruction by the body's immune system. These abilities highlight the intricate nature of cancer, which affects individual cells and disrupts the body's overall biological systems.

The influence of cancer goes beyond the cellular level, impacting tissues, organs, and complete biological systems. It disrupts the regular operation of body systems like circulation, lymphatic, and immune systems, playing a part in the systemic characteristics of the illness. This complex interruption highlights the essential need for a holistic strategy in cancer research and treatment, considering the complex interaction of genetic, molecular, cellular, and systemic elements. As we explore further into the biological systems impacted by cancer, it is clear that comprehending and addressing this illness necessitates a comprehensive perspective that incorporates information from various aspects

of the biological system.

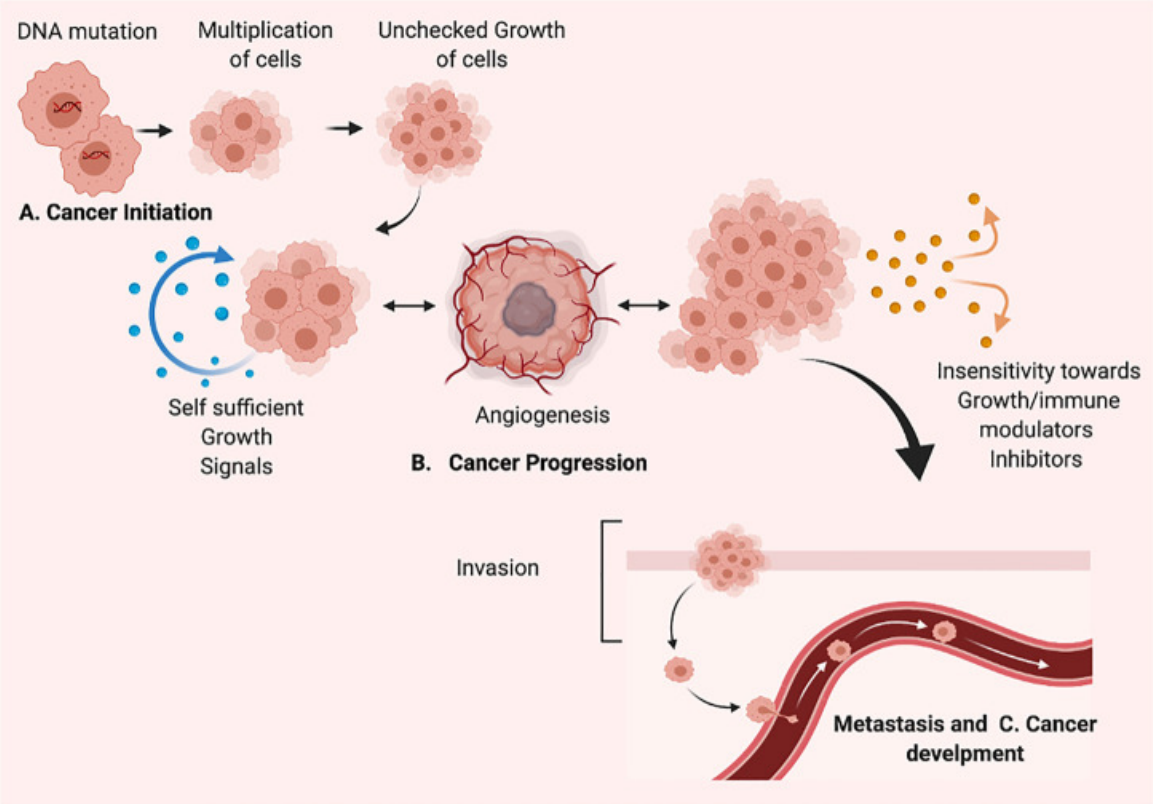


Figure 1.1: ***Figure [45]*** *The process of cancer development/carcinogenesis, (A) Cancer initiation (B). Cancer Progression (C). Metastasis and Cancer development.*

1.1.2 . Multiple Levels of the Biological System

Cancer's impact on the body extends from DNA-level changes to visible symptoms in patients [111]. This complex series of events, from genetic mutations to tissue damage, provides a clearer understanding of the disease's progression and manifestations.

The DNA is where cancer develops, with mutations changing the genetic makeup. These genetic changes can switch off tumor suppressor genes or turn on oncogenes, leading to uncontrolled cell growth seen in cancer. DNA mutations are not independent; instead, they increase complexity at the RNA level. In this case, the transcription process may increase these genetic abnormalities, as mRNA molecules contain mutated guidelines that interfere with typical gene expression patterns and cellular operations [250].

As cancer progresses, the impact becomes stronger as the modified genetic material is used to

make proteins. Proteins, responsible for carrying out numerous cellular tasks, function according to incorrect guidance, leading to errors in critical signaling pathways and cellular activities [194]. A clear example of this misdirection in cancer cells is seen as abnormal protein activity leading to constant cell multiplication and infiltration of neighboring tissues.

The metabolic landscape extends beyond proteins and is altered by cancer to fulfill the needs of rapid cell growth. This change in metabolism is a hallmark of cancer, not only promoting tumor growth but also creating a unique environment for the tumor [42]. The diverse changes in metabolites play a crucial role in the survival and proliferation of cancer cells, reinforcing the presence of the disease.

The combined effects of these molecular and metabolic disturbances are noticeable on a cellular level. Cancer cells come together to form tumors as they acquire the capability to multiply continuously and avoid death. However, these tumors do not exist alone. They engage with and change the structure of the nearby tissue, disturbing regular organ operation and setting off a series of harmful consequences [11].

This disruption at the tissue level causes cancer to affect not only the tumor but the entire organism, resulting in clinical symptoms that affect the quality of life of the patient and may eventually lead to the patient's death. The impact of cancer on individual patients, which includes symptoms like fatigue and pain, as well as organ dysfunction, highlights the disease's ability to affect the body's systems and overall well-being.

This progression is done from DNA to the patient, with each level closely linked and impacting the succeeding one. This exploration of the biological stages of cancer highlights the importance of taking a thorough approach in research and treatment, which recognizes the complex nature of the disease and aims to intervene at various stages of its progression. By comprehending and focusing on the various effects of cancer, we get closer to creating treatments that are just as adaptable and complex as the disease, providing optimism for improved and individualized therapies.

1.1.3 . Precision medicine

Precision medicine represents a revolutionary evolution in healthcare, especially in oncology, as personalized treatments can provide better care and enhance patient results [5]. Precision medicine seeks to personalize healthcare by utilizing information about a person's genetics, lifestyle, and environment to tailor treatments and prevention strategies to their needs. This method differs significantly from the traditional one-size-fits-all approach, providing a more detailed insight into disease

processes and treatment outcomes.

The rapid advancement of sequencing technologies is driving the growth of the precision medicine field. These advancements have opened up new possibilities by offering vast data in various areas such as genomics, transcriptomics, proteomics, and metabolomics, in addition to conventional clinical information [180]. Sequencing a person's genome rapidly and affordably is now a vital part of precision medicine, allowing us to pinpoint genetic mutations linked to different types of cancer. This abundance of diverse omic data provides a comprehensive molecular perspective on the patient's illness, establishing the foundation for individualized treatment plans when combined with comprehensive patient health records and histopathological images.

Nevertheless, the challenges of managing complex and high-dimensional data come with noteworthy obstacles despite the opportunities brought forth by technological advancements [167]. The extensive data produced by advanced sequencing and other omic technologies have positives and negatives, providing critical insights into cancer's molecular basis yet creating difficulties in storing, managing, analyzing, and interpreting the data. Novel methods are required to effectively analyze the massive amount of complex data that surpasses the capabilities of traditional tools and techniques.

In this context, AI offers new crucial tools and methods to enhance precision medicine in oncology[126]. AI, utilizing its sophisticated machine learning algorithms and deep learning frameworks, provides the computational power and expertise required to analyze the intricacies of multi-omics data. AI algorithms are very good at recognizing patterns and connections in large datasets that would be impossible for humans to analyze. AI-powered tools are leading advancements in cancer care, from predicting disease risk using genetic information to finding potential treatment targets and predicting how patients will respond to therapies. They assist in examining high-dimensional data and incorporate various data types, providing a more thorough comprehension of cancer's molecular landscape and its relationship with environmental and lifestyle factors.

The move towards AI-powered precision medicine in cancer care mirrors a more significant change in healthcare, highlighting the importance of sophisticated computational resources for handling and analyzing the intricate nature of contemporary biomedical information. Advancing, the collaboration of AI and precision medicine offers the potential to discover uncharted territories in cancer treatment, bringing personalized, effective, and efficient care to patients globally. This fusion of fields marks a forthcoming era in which extensive data and advanced computing capabilities come together, open-

ing pathways for innovations that seek to revolutionize cancer treatment.

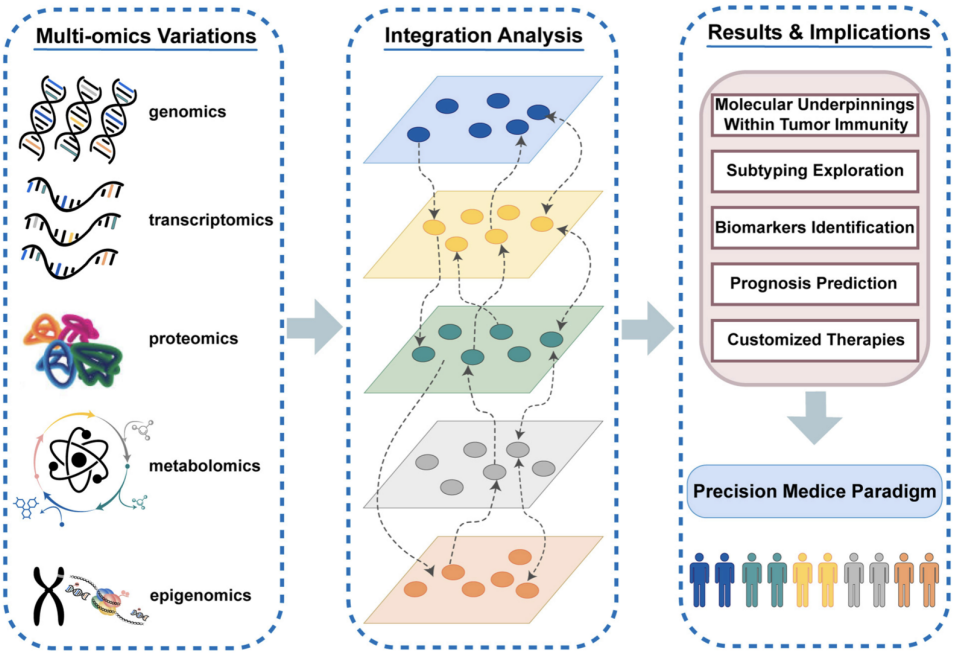


Figure 1.2: ***Figure extracted from [52]*** *The interactive multi-molecular layer systems and the vital role of multi-omics variations in tumor immunity and immunotherapy.*

1.2 . Multiple Modalities for Precision Medicine

1.2.1 . Multi-Omics Data

Multi-omics provides a powerful approach to biological research and medical diagnostics by integrating various biological data types to understand an organism's functions comprehensively. This approach comprises several "omics" disciplines: genomics focuses on DNA; transcriptomics analyzes messenger RNA molecules; proteomics examines proteins, including their expression and functions; and epigenomics studies epigenetic changes in the genetic material of cells.

Genomic Data

Genomic data encompasses the comprehensive study of an organism's DNA, including its genes. In cancer research, genomic data provides crucial insights into the genetic alterations driving tumor development, progression, and response to treatment. By analyzing genomic sequences, researchers can identify mutations, structural variations, and other genetic abnormalities contributing to the com-

plex landscape of cancer biology. The advent of high-throughput sequencing technologies has revolutionized the field, enabling large-scale projects like The Cancer Genome Atlas (TCGA) to generate vast amounts of genomic data across diverse cancer types [90, 197, 183].

Mutation data refers to identifying and characterizing changes in the DNA sequence that can lead to cancer. These mutations can be categorized as somatic or germline, with somatic mutations occurring in the DNA of individual cells during a person's lifetime and germline mutations being inherited. High-throughput sequencing technologies, such as whole-genome sequencing (WGS) and whole-exome sequencing (WES), are employed to detect these mutations. WGS provides a comprehensive view of all genetic alterations across the entire genome, while WES focuses on the coding regions of the genome, where most disease-causing mutations occur. By analyzing mutation data, we can pinpoint driver mutations that play a critical role in tumorigenesis and identify potential therapeutic targets [90, 187].

Copy number variations (CNVs) are a type of structural genetic alteration where genome segments are duplicated or deleted, leading to changes in the number of copies of particular genes. CNVs can significantly impact gene expression and contribute to cancer development and progression. Technologies such as array comparative genomic hybridization (aCGH) and next-generation sequencing (NGS) are commonly used to detect CNVs. aCGH involves comparing the DNA of cancer cells with normal cells to identify regions of genomic gain or loss. At the same time, NGS provides a high-resolution view of CNVs across the genome. Analyzing CNV data helps researchers understand the genomic instability of tumors, identify oncogenes and tumor suppressor genes affected by these variations, and develop targeted therapies to address these genetic abnormalities [90, 203].

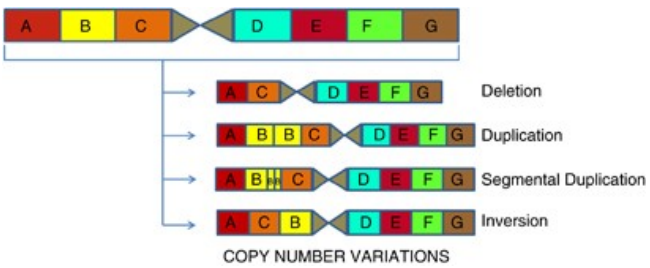


Figure 1.3: ***Figure extracted from [6]*** *Types of Genomic Variants: Genomic variants, such as CNVs, can be categorized into deletions, duplications, segmental duplications, and inversions. These variations may affect an entire gene or just a portion of a gene, as illustrated in the figure.*

By integrating mutation and CNV data, we can gain a comprehensive understanding of the genetic

landscape of cancer, enabling the identification of novel biomarkers and the development of precision medicine strategies tailored to individual patient's genomic profiles [90, 41].

Transcriptomic Data

Transcriptomic data characterize the RNA transcripts produced by the genome under specific circumstances or in specific cell types. This data provides insights into gene expression patterns, regulatory mechanisms, and the functional state of cells. In cancer research, transcriptomics is essential for understanding how changes in gene expression contribute to tumor development, progression, and response to treatment. High-throughput technologies like RNA sequencing (RNA-seq) have significantly advanced the field, allowing researchers to capture the complexity of the transcriptome in unprecedented detail [252, 106, 183].

Transcriptomic data can be collected at different scales: bulk RNA-seq captures average gene expression across many cells, single-cell RNA-seq (scRNA-seq) reveals gene expression in individual cells, and spatial transcriptomics provides gene expression data within the spatial context of the tissue. Each scale offers unique insights, enhancing our understanding of gene expression dynamics in cancer.

One type of transcriptomics data is RNA sequencing (RNA-seq). It is a powerful technique to analyze a sample's complete RNA transcripts, including coding and non-coding RNAs. RNA-seq provides a quantitative and qualitative snapshot of gene expression, enabling the identification of differentially expressed genes between normal and cancerous tissues. This technology involves converting RNA into complementary DNA (cDNA), sequenced using high-throughput sequencing platforms. RNA-seq data allows researchers to detect gene fusions, alternative splicing events, and novel transcripts, offering a comprehensive view of the transcriptomic alterations in cancer. By analyzing RNA-seq data, researchers can uncover vital regulatory pathways, identify potential therapeutic targets, and understand the molecular mechanisms driving cancer [252, 179].

Another type is microRNAs (miRNAs). Small, non-coding RNAs play critical roles in post-transcriptionally regulating gene expression. miRNAs bind to messenger RNAs (mRNAs) and either degrade them or inhibit their translation, thereby controlling the expression of target genes. In cancer, dysregulation of miRNAs can contribute to tumorigenesis, metastasis, and resistance to therapy. Techniques such as miRNA sequencing (miRNA-seq) and microarray profiling are used to study miRNA expression patterns. miRNA-seq involves sequencing the small RNA fraction of a sample to identify and quantify

miRNAs, while microarray profiling uses hybridization techniques to measure miRNA levels. Analyzing miRNA data provides insights into the regulatory networks that influence cancer progression and offers potential biomarkers for diagnosis, prognosis, and therapeutic intervention [19, 114, 88].

By integrating RNA-seq and miRNA data, we can gain a holistic understanding of the transcriptomic landscape of cancer. This integrative approach allows for the identification of complex regulatory interactions between coding and non-coding RNAs, enhancing our ability to develop precision medicine strategies that target specific molecular pathways and improve patient outcomes [252, 106, 88].

Epigenomic Data

Epigenetic changes like DNA methylation are vital in regulating gene activity without altering the genetic code. DNA methylation involves adding a methyl group to cytosine bases, primarily at CpG dinucleotides, and can either activate or silence genes by affecting transcriptional access [21, 128].

In cancer, abnormal methylation patterns, such as tumor suppressor genes' hypermethylation or oncogenic pathways' hypomethylation, contribute to tumor development and progression. These reversible changes in gene regulation are promising targets for cancer detection and treatment [21, 73].

Advances in bisulfite sequencing and array-based methods have made it possible to analyze methylation patterns across entire cancer genomes, leading to the identification of epigenetic markers for early diagnosis, prognosis, and personalized treatment [143, 28].

Integrating epigenetic data with genomic, transcriptomic, and proteomic information enhances our understanding of the molecular interactions in cancer, paving the way for precision medicine approaches that target specific epigenetic modifications to restore normal gene function [4, 20].

However, challenges remain in interpreting DNA methylation data, as patterns can vary across tissues and environmental contexts. Establishing transparent cause-and-effect relationships between methylation changes and cancer characteristics requires robust experimental designs and long-term studies to understand the role of epigenetics in cancer fully [78, 217].

1.2.2 . Histopathology Slides

Examining tissue at a microscopic level to observe disease manifestations, known as histopathology, is a crucial component of diagnostic pathology, particularly in oncology. This careful practice requires a thorough examination of tissue specimens, usually obtained through biopsies or surgery,

allowing oncologists to evaluate the existence, spread, and type of neoplastic (cancerous) conditions. Histopathological evaluation allows pathologists to give essential diagnostics, provide prognostic insights, and offer guidance for strategic therapeutic decisions, making them essential in cancer patient care.

The fundamental aspect of histopathology involves directly observing and examining cellular structure and tissue organization through a microscope. These observations are crucial for various purposes: diagnosis and classification, grading and staging, assessment of tumor margins, evaluation of cell differentiation, or detection of molecular biomarkers.

Understanding the distinct features of cancer on histopathology slides can help tailor treatments to match the individual patient's disease profile more effectively, thus enhancing effectiveness while reducing unwanted side effects. For example, if hormone receptors are not present, hormone therapy may not be the best option, but if a patient has high levels of HER2/neu protein, they may benefit from HER2-targeted therapies.

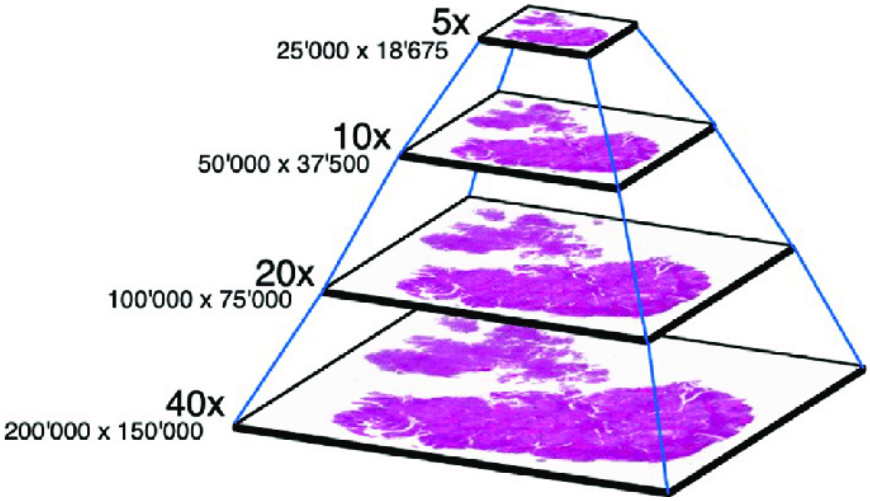


Figure 1.4: ***Figure [166]*** *An example of WSI format including multiple magnification levels. The size of each image of the pyramid is reported under the magnification level in terms of pixels.*

Clinical Extraction of Histopathological Slides

Extracting and digitalizing histopathological slides is essential in transitioning from traditional microscopy to digital pathology, enabling enhanced diagnostics, collaboration, and analytical capabilities [189, 162].

Tissue samples are typically obtained through surgical excisions or biopsy procedures, ensuring

sufficient material for comprehensive evaluation while maintaining patient safety. Post-collection, samples are fixed, usually with formalin, to preserve cellular and molecular integrity. They are then embedded in paraffin wax, facilitating the sectioning of thin tissue layers necessary for microscopic analysis [142, 98].

Thin tissue sections are prepared using a microtome, placed on slides, and stained to enhance visibility. Staining techniques in histopathology are essential for enhancing the visualization of cellular and tissue structures, facilitating the diagnosis and understanding of various diseases. These methods selectively color different components of tissue specimens, allowing pathologists to distinguish between cellular elements and identify pathological changes [200].

Hematoxylin and Eosin (H&E) staining is the most widely used technique due to its effectiveness in delineating tissue morphology. Hematoxylin imparts a blue hue to cell nuclei, emphasizing DNA and RNA contents, while eosin stains the cytoplasm and extracellular matrix in shades of pink and red. This contrast is invaluable for routine histological examination and is foundational in pathology [81].

Immunohistochemistry (IHC) provides a more targeted approach by detecting specific antigens in the tissues using antibodies. This method is crucial in oncology for identifying protein expressions, such as hormone receptors in breast cancer, that directly influence treatment decisions. IHC employs chromogenic labels or fluorescent tags that bind to antibodies, making the antigen-antibody reaction visible under a microscope [135, 91].

Other special stains address the need to highlight particular tissue components not adequately displayed by H&E. For instance, the periodic acid-schiff (PAS) stain detects polysaccharides, coloring them magenta, and helps identify fungal organisms and glycogen deposits. Masson's Trichrome stain differentiates between muscle (stained red), collagen (blue), and cytoplasm (pink or light red), commonly applied in connective tissue disease assessments. Another proper stain, Giemsa, enhances the visibility of blood cells and is extensively used in hematology and for detecting parasites [206, 222].

Whole-Slide Images

Whole slide images (WSIs) are digital versions of complete biopsy slides with a multi-resolution, pyramidal structure for viewing at different magnifications, replicating the traditional microscope experience. This structure is pyramid-shaped and includes tiers, each representing the same tissue space at varying levels of detail. This enables pathologists to quickly move through the entire slide at

a lower level of detail and then focus on specific areas at higher levels for thorough analysis. Standard magnifications vary from 5x to 40x, with premium scans at 20x and 40x offering the precision essential for precise cellular examination. The extensive scans produce big files, usually between 100 MB and 1 GB, depending on resolution and tissue complexity, requiring ample storage and robust IT systems [76, 140].

The significant sizes of WSIs present particular obstacles, particularly when applying deep learning techniques for image examination. Training deep learning models on such huge images demands substantial memory and computational resources. The limited RAM in typical computer systems hinders the ability to load several WSIs at once, making it challenging to train models effectively with large batches of data. In order to address these problems, researchers frequently use patch-based training methods, breaking down every image into smaller, more manageable sections. This approach decreases the memory burden by enabling consecutive training on image patches, though it requires advanced methods to uphold the image's contextual coherence. Furthermore, reducing resolution through down-sampling methods is employed in training, which may lead to missing important diagnostic information at times [102, 38].

Addressing these challenges is essential for successfully implementing machine learning in digital pathology. Continuing research focuses on improving data processing and training algorithms that can handle large image datasets while preserving relevant details. These developments are crucial for maximizing the diagnostic capabilities of WSIs in pathology, improving precision, and facilitating the progression of personalized medicine [68, 35].

Regions of Interest

Histopathology slides contain a variety of cellular structures and complex tissue architectures, each offering important information about diseases. Identifying and analyzing certain areas of interest (ROIs) in these slides is essential for diagnosing and comprehending pathological conditions. Below is a summary of the main ROIs found in histopathology slides, including their descriptions and importance in medical diagnosis.

The tumor region is typically the primary focus in oncologic histopathology. This area contains the cancerous cells and is critical for determining the type of cancer, its aggressiveness, and potential response to treatment [177]. Pathologists look for features such as the size and shape of the tumor cells, the pattern of their arrangement, and the presence of necrosis or mitotic figures to assess tumor

grade and stage [140].

Stromal regions refer to the supportive tissue around cancer cells, composed mainly of connective tissues and blood vessels. The stromal region is essential for understanding the tumor microenvironment, including how cancer interacts with surrounding tissues and the extent of angiogenesis (formation of new blood vessels), which is often a prerequisite for tumor growth and metastasis [267]. Changes in the stroma can indicate tumor invasion and aggressiveness [186].

Inflammatory regions, characterized by a high concentration of immune cells, are crucial in conditions such as chronic inflammation and autoimmune diseases. In cancer, these regions reflect the immune system's response to the tumor [122]. Assessing the type and level of immune cell infiltration can offer valuable information about the prognosis and help guide immunotherapy treatments [59].

Necrotic regions within a tumor indicate areas where cancer cells have died, often due to insufficient blood supply. The presence and extent of necrotic tissue can be a marker of tumor progression and is generally associated with a more aggressive disease state [193]. Necrotic regions are also important in treatment planning, especially in predicting the response to radiation therapy [130].

Marginal regions are the boundaries between the tumor and normal tissues. Marginal regions are critical in surgical pathology to ensure that the surgical resection margins are free of cancer cells, which is essential for reducing the risk of recurrence [242]. The characteristics of the tumor at these margins can provide information about its invasiveness and the likelihood of complete surgical removal [268].

Lymphoid structures, such as lymph nodes, often appear in histopathology slides, especially in cancer cases where they are assessed for metastasis. Cancer cells in lymph nodes are a critical factor in cancer staging, directly impacting treatment decisions and prognosis [205].

Each of these regions of interest provides specific information contributing to the overall disease diagnosis and management. In oncology, understanding these regions' distinct roles and features helps pathologists provide detailed and accurate reports that guide effective treatment strategies. As digital pathology and image analysis technology advance, the precision in identifying and interpreting these regions continues to improve, offering deeper insights and enhancing personalized medicine approaches in cancer care [162].

A diversity of cell types

In histopathology, tissue samples are examined under a microscope to identify and characterize different cell types. Tumor or cancer cells are a primary focus due to their abnormal behavior, including uncontrolled growth and resistance to cell death [106, 87]. These cells are identified by features like enlarged nuclei, increased mitosis, and disorganized tissue structure [140]. Immunohistochemical staining and molecular profiling provide further insights into these cells' characteristics and potential treatment targets [135].

Stratified epithelial cells, organized in layers, form protective tissues in areas subjected to mechanical stress, such as skin and the esophagus. Histopathological assessment of these cells focuses on the structure and integrity of each layer, which is critical for diagnosing conditions like squamous cell carcinoma [81, 165].

Necrotic cells, which result from tissue injury, are identified by features like cellular swelling and membrane rupture [72]. Their presence indicates tissue damage and inflammation, making them crucial for assessing the extent of the injury and guiding treatment [219].

Connective tissue cells, including fibroblasts, adipocytes, chondrocytes, and osteocytes, maintain the tissue's structural and functional integrity. Histopathological evaluation of these cells aids in diagnosing disorders like fibrosis and skeletal abnormalities [139, 208, 198].

Tumor-infiltrating lymphocytes (TILs) are immune cells within tumors that can influence tumor growth and response to therapy [84]. Their presence and activity are critical indicators of prognosis and can guide immunotherapy decisions [59, 238].

1.3 . Introduction to Artificial Intelligence & Statistical Learning

Artificial Intelligence (AI), a crucial advancement in contemporary technology, is transforming different industries, such as healthcare, by being able to carry out tasks that usually demand human intelligence [62]. At its core, artificial intelligence allows machines to learn from information, make choices, and resolve problems by imitating cognitive abilities like reasoning, learning, and comprehending language. AI is especially significant for healthcare providers and doctors in diagnostics, personalized medicine, and drug discovery.

In diagnostics, AI improves the precision and effectiveness of analyzing medical images [195]. By

utilizing sophisticated algorithms, AI technology can rapidly examine X-rays, MRIs, and CT scans, recognizing patterns that suggest illnesses like cancer or neurological disorders more quickly and precisely than conventional techniques. This ability assists radiologists not only in detecting conditions early but also in handling their heavy workload.

AI's impact on personalized medicine is just as significant. AI models can utilize genetic information and detailed clinical data to forecast how individuals respond to treatments. This method is particularly advantageous in oncology, as AI-powered information can guide customized treatment strategies that enhance patient results and reduce adverse effects.

Moreover, AI speeds up the process of drug discovery and development, which has historically been both lengthy and expensive [7]. AI tools can predict the effectiveness of chemical compounds, simulating their impacts to identify potential drug candidates sooner in the process, from lab research to clinical trials, making the journey more efficient.

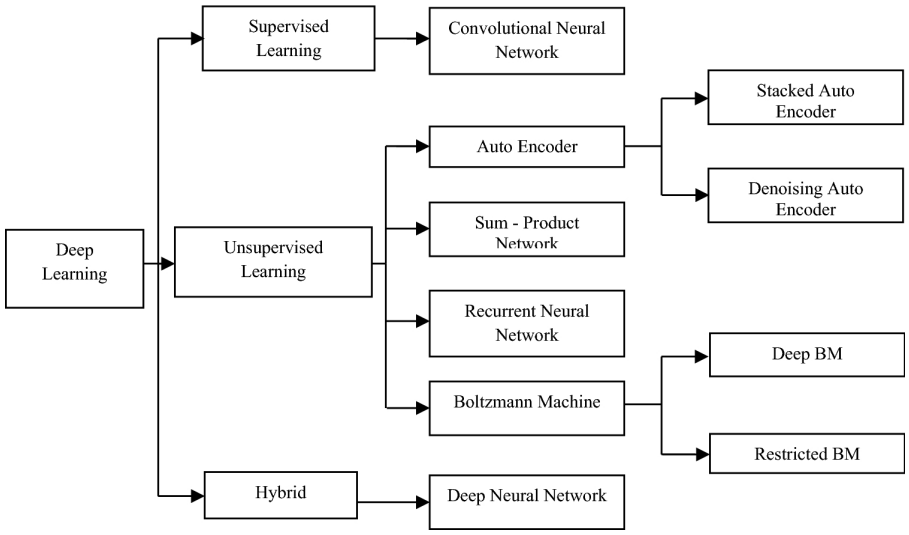


Figure 1.5: ***Figure [164]*** *Proposed Classification model of Deep Learning by Manimaran et al.*

1.3.1 . General Idea

AI and statistical learning are fundamental to modern biostatistics, enabling data interpretation, prediction, and decision-making by uncovering patterns within complex datasets.

Basic statistical models, such as linear and logistic regression, are crucial tools in analyzing data. Linear regression predicts continuous outcomes based on one or more predictors, while logistic regression is used for binary outcomes [83]. The Cox proportional hazards model is essential in survival

analysis, providing insights into treatment effects on time-to-event data. However, these models require specific assumptions about the data structure, which can limit their flexibility in exploratory analysis and hypothesis testing.

As the complexity of data increases, machine learning (ML) methods have become powerful alternatives to traditional statistical approaches. Techniques like decision trees, random forests, and support vector machines excel at extracting patterns and making predictions from large datasets [29]. These methods are precious in biomedicine, where they are used for disease diagnosis, predicting patient outcomes, and identifying new therapeutic targets from genetic, proteomic, and clinical data.

Deep learning further enhances the capabilities of ML by utilizing neural networks with multiple layers, enabling the analysis of intricate data patterns. Convolutional Neural Networks (CNNs) are particularly effective in image analysis, as they detect spatial features like edges and textures, making them ideal for tasks such as tumor detection and tissue classification [101]. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are designed to handle sequential data, maintaining information across time steps, which is essential for applications like language modeling and time-series forecasting [99, 243]. Meanwhile, Transformers, which use attention mechanisms to process sequential data and capture long-range dependencies, have driven significant advancements in natural language processing and genomic data analysis, exemplified by models like GPT and BERT [246, 263, 65].

Foundation models represent a significant leap forward, as they are trained on diverse datasets, providing a comprehensive understanding of various data types. These models can be fine-tuned for specific biomedical applications, enhancing personalized medicine, drug discovery, and patient care by enabling detailed analysis of complex biological information.

1.3.2 . A myriad of tasks

In artificial intelligence (AI) and machine learning, data analysis techniques are categorized into unsupervised, supervised, and self-supervised learning, each with specific applications in the biomedical field.

Unsupervised learning identifies patterns and structures in data without labeled outcomes. It is particularly effective for clustering and dimensionality reduction, making it valuable in biomedicine for analyzing large genomic datasets to uncover new disease subtypes and potential therapeutic targets

[173, 257, 234].

Supervised learning, on the other hand, uses labeled data to train models to make predictions on new inputs. This approach is crucial in diagnostic imaging, where algorithms are trained on labeled images to detect pathologies in new scans. It is also used in predictive modeling, where patient data, including clinical and molecular biomarkers, is analyzed to forecast disease progression or treatment response, thereby improving diagnostic accuracy and optimizing therapeutic strategies [153, 75, 239]. Representation learning, an essential aspect of supervised learning, enables the model to automatically discover the features needed for these tasks, further enhancing its predictive capabilities.

Self-supervised learning bridges the gap between unsupervised and supervised methods by generating labels from the data. This approach is auspicious in biomedicine, where large amounts of unlabeled data, such as genomic sequences or patient records, can be used to discover meaningful patterns. Self-supervised learning can accelerate the identification of biomarkers or therapeutic targets, even without extensively annotated datasets [125, 15]. Additionally, it supports multimodal integration, where different types of biomedical data—such as imaging, genomics, and clinical records—are combined to provide a more comprehensive understanding of complex biological processes and improve the precision of predictive models.

1.3.3 . Application to Biomedical Sciences

Incorporating AI into biomedicine is fundamentally changing how medical issues are addressed, from general clinical applications to the specific challenges of cancer treatment. Researchers and clinicians leverage AI to significantly advance diagnosis, treatment, and personalized healthcare through supervised, unsupervised, and self-supervised learning techniques. This integration of AI is enhancing medical procedures and paving the way for new approaches in patient care, particularly in the realm of precision medicine [124, 239].

In general medicine, AI plays a vital role in diagnostic imaging, patient monitoring, and the analysis of electronic health records. AI algorithms are now indispensable in accurately diagnosing diseases by analyzing medical images such as X-rays, MRIs, and CT scans [153]. Beyond diagnostics, AI is used in predictive analytics to forecast patient outcomes based on clinical data, thereby improving treatment plans through insights derived from past patient data [75, 185].

In oncology, AI transforms cancer treatment by identifying genetic mutations and biomarkers associated with various cancers. This aids in early detection and the development of precise treatments

through the analysis of large genomic datasets [218, 169, 214]. AI also personalizes chemotherapy plans, predicts patient responses to treatments, and monitors disease progression, leading to better outcomes [141]. Moreover, AI accelerates the discovery of new therapeutic targets and streamlines drug development by predicting the interactions between drugs and biological pathways, thereby reducing the time and cost associated with traditional drug development methods [251, 47].

A crucial aspect of AI in biomedicine is its application in survival analysis, particularly in oncology. Traditional models like the Cox proportional hazards model have been essential for predicting patient survival and treatment outcomes. However, advancements like DeepSurv and Cox-net have enhanced these predictions by incorporating deep learning to manage complex, non-linear relationships in survival data. These models provide more accurate survival predictions and can integrate high-dimensional data, making them especially valuable in omics data analysis, further refining personalized treatment strategies.

The intersection of AI and precision medicine leads to highly individualized medical care based on the unique characteristics of each patient. AI offers the computational power to analyze diverse data types—including genomics, transcriptomics, proteomics, histopathology images, and clinical records—allowing for the identification of distinct disease patterns and the prediction of the most effective treatments for individual patients [148]. This approach is particularly beneficial in cancer care, where AI-driven analysis can match treatments to a patient's specific tumor profile, minimizing side effects and improving overall outcomes [239, 144].

1.3.4 . The Clinical Interpretability Challenge

AI revolutionizes clinical practice by enhancing diagnostics, personalizing treatments, and improving patient outcomes. AI systems are particularly adept at analyzing complex data, such as medical images and genetic sequences, making them indispensable tools for clinicians. However, a significant challenge is the interpretability of these AI systems, especially those using deep learning, which often operate as "black boxes" [236]. This lack of transparency can undermine trust, as clinicians must understand the reasoning behind AI-generated decisions to ensure they align with medical standards.

To address this, explainable AI (XAI) techniques are being developed to make AI models more transparent. These techniques help clarify the data features that influence AI decisions, bridging the gap between complex algorithms and clinical expertise.

Interpretability refers to the degree to which a human can understand the internal workings of

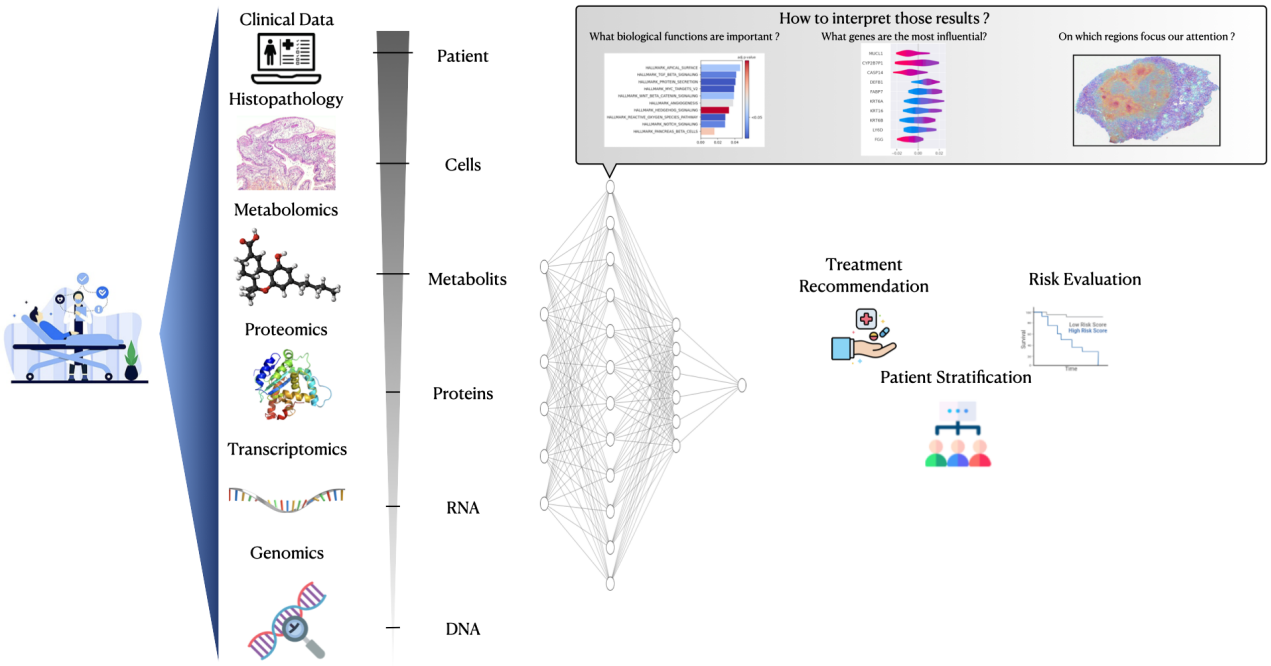


Figure 1.6: *An overview of multimodal integration for precision medicine*

an AI model and the rationale behind its decisions. This is essential in medical contexts where clinicians must grasp how AI systems arrive at their predictions to ensure trust and informed decision-making. On the other hand, explainability focuses on making the AI model outputs—its decisions and predictions—clear and understandable to humans. This is often achieved through visualizations or feature importance scores highlighting which factors influenced the model's decision [89]. While interpretability deals with the transparency of the model's inner processes, explainability ensures that the end user can comprehend the reasoning behind specific outcomes. These concepts are crucial for integrating AI into healthcare while maintaining ethical and patient-centered care.

Explainability methods can be categorized as model-agnostic or model-specific. Model-agnostic methods, like LIME and SHAP, can be applied to any AI model to explain its predictions. In contrast, model-specific methods are tailored to specific models, such as saliency maps for deep learning or attention mechanisms in transformers. Additionally, explainability can be global, providing insights into the overall model behavior, or local, focusing on individual predictions. Post-hoc methods explain decisions after the model is trained, whereas intrinsic methods involve naturally transparent models, such as decision trees.

Balancing AI's advanced analytical capabilities with the need for interpretability is essential to fully leverage AI in clinical practice, ensuring that these technologies enhance patient care without sacrificing clarity and trust.

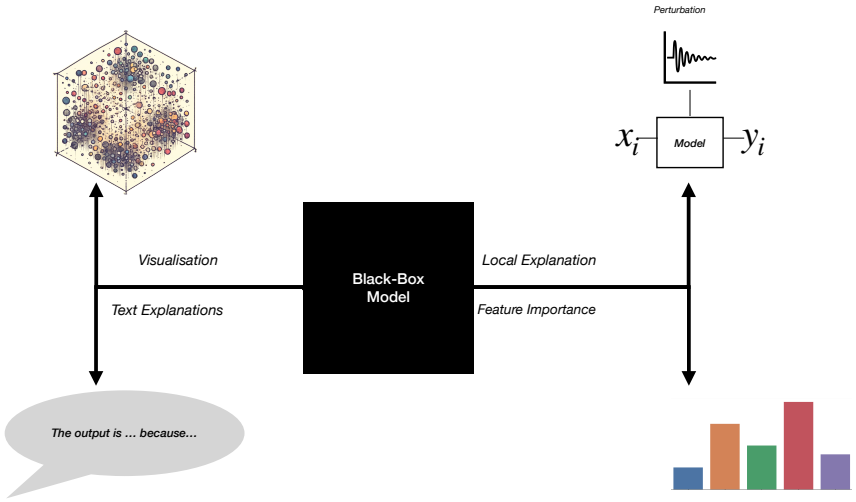


Figure 1.7: *Forms of Explainability*

1.4 . Contributions

This thesis adds to the developing field of precision medicine by introducing innovative approaches and models designed to improve the incorporation and understandability of intricate biomedical data. This work aims to tackle the crucial issue of integrating and analyzing data from various omics sources and histopathology to understand cancer mechanisms better. The thesis's contributions can be outlined in the following way:

Firstly, we present novel approaches for effectively combining information from various omic sources using an understandable artificial intelligence framework. Our approach combines various types of data from genomics and transcriptomics to create a comprehensive and multifaceted molecular depiction of cancer, acknowledging the abundant yet intricate information in these fields. Through the utilization and development of explainable AI methods, the incorporation procedure helps improve our models' predictive capabilities and precision. It maintains clarity and comprehensibility for clinicians and researchers. This combination of integration and explainability helps bridge a significant gap in current methods, allowing for a better grasp of biological processes and building confi-

dence in AI-based findings.

Additionally, the thesis contributes to the digital histopathology field by creating innovative approaches to designing histopathology slide representations. These techniques focus on integrating crucial details, like spatial connections and interactions in the tumor microenvironment, into the examination. Acknowledging the significance of these elements in grasping tumor behavior and patient prognosis, our strategy utilizes advanced image analysis and deep learning methods to extract and interpret these essential data points. In addition to these advances, we are launching new explainability tools designed explicitly for histopathology analysis. These tools should help users navigate and comprehend the intricate decision-making processes required for diagnosing and characterizing tumors from slide images. This advancement is a significant step towards incorporating and understanding histopathology better in cancer research and treatment planning.

Ultimately, we introduce an all-encompassing structure to combine omics data and histopathology analyses in a transparent AI system. This structure is created to provide various levels of comprehensibility, matching the diverse levels of the biological system, including molecular interactions, cellular processes, tissue organization, and overall patient health. In this way, it meets an essential requirement in precision medicine for tools that can efficiently integrate different data types and offer an understanding of how they are interpreted across multiple biological levels. This comprehensive method guarantees that the analyses produced are technically robust and closely linked to biological things, improving the effectiveness of AI-based techniques in clinical decision-making and customized treatment.

Collectively, these contributions highlight the capability of AI and machine learning to revolutionize how we conduct cancer research and treatment. This thesis sets the foundation for future progress in precision medicine by integrating multi-omics data and histopathology and creating an explainable framework. It aims to improve cancer care by making it more personalized, effective, and understandable.

Multimodal Representation Learning for High-Dimensional Data

Contents

Abstract

This chapter provides a comprehensive state-of-the-art review of representation learning techniques and benchmarks current multimodal integration strategies for multi-omics data. By evaluating essential methods such as feature-level fusion, decision-level fusion, and intermediate integration, we assess their effectiveness in capturing the relationships and complementarities between different data types. Our comparative analysis highlights the strengths and weaknesses of each approach, emphasizing the need for improved integration strategies. This foundational assessment sets the stage for developing novel strategies, which will be the focus of the subsequent chapters.

Representation learning is a fundamental concept in machine learning. It focuses on creating algorithms that can autonomously identify the necessary representations for detecting features or classifying data from its raw form. This method is essential in various machine learning domains, such as deep learning, as it allows a machine to recognize and best utilize the inherent patterns in the data. We will show how representation learning is essential in precision medicine within the field of oncology as it helps combine various forms of data, like multi-omics and histopathology data, to improve a patient's prognosis.

The primary goal of representation learning is to convert raw data into a more understandable format by further processing layers or machine learning models. This transformation is intended to reveal significant characteristics frequently concealed in the raw data, rendering the data more suitable for analysis and decision-making procedures. This could involve transforming detailed gene expression profiles or complex patterns in histopathology images into more relevant characteristics from a biological or clinical perspective.

2.1 . Representation Learning

Representation learning primarily focuses on identifying an appropriate series of processing steps for input data, often utilizing a mix of linear and non-linear transformations. Representation learning techniques can be divided into supervised, unsupervised, and semi-supervised methods.

2.1.1 . Supervised Representation Learning

Supervised learning methods focus on effectively learning from labeled data to map inputs to outputs.

Deep learning models, particularly neural networks like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are standard in this category. They learn hierarchical representations through layers, where each layer captures increasingly abstract data features. The mathematical model can be represented as:

$$\phi(x) = h^{(n)}(\dots h^{(2)}(h^{(1)}(x)))$$

where $h^{(i)}(x) = \sigma(W_i x + b_i)$, σ is a non-linear activation, and W_i, b_i are layer parameters [144, 216, 93].

2.1.2 . Unsupervised Representation Learning

Unsupervised methods focus on understanding the structure of unlabeled data and identifying patterns without reference to known outcomes.

Principal Component Analysis (PCA) reduces dimensionality by finding orthogonal projection directions that maximize variance:

$$W^* = \arg\max_{W \in \mathbb{R}^{d \times k}} \{\text{Tr}(W^T X^T X W)\} \quad \text{subject to} \quad W^T W = I$$

This technique is effective for initial data exploration and simplifying complex datasets with minimal loss of information [127].

Autoencoders

In the same vein, we have autoencoders. It is an architecture that can be seen as a non-linear extension of PCA. We consider our data to be a matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ and we want to represent it as a matrix $\mathbf{Z} \in \mathbb{R}^{N \times d}$ with $d < D$. The autoencoder is composed of two parts, an encoder function $q : \mathbb{R}^{N \times D} \to \mathbb{R}^{N \times d}, X \to q(\mathbf{X}) = \mathbf{Z}$ and a decoder function $p : \mathbb{R}^{N \times d} \to \mathbb{R}^{N \times D}, Z \to p(\mathbf{Z}) = \hat{\mathbf{X}}$ [116].

The encoder and decoder function are a sequence of multiple non-linear layers of the form $f(u) = \sigma(\mathbf{W}u + b)$ where u is the layer input, $\mathbf{W}$ is the weight matrix,b is the bias and σ is called an activation function. This means that both the encoder and decoder have a set of parameters such that $\mathbf{Z} = q(X; \theta_e)$ and $\hat{\mathbf{X}} = p(\mathbf{Z}; \theta_d)$ [249].

The goal of the architecture is to optimize θ_e and θ_d in order to minimize the reconstruction error through a reconstruction loss function:

$$\mathcal{L}_{recon}(\mathbf{X};\theta_e,\theta_d) = \frac{1}{N}||X-\hat{\mathbf{X}}||_2^2 = \frac{1}{N}||\mathbf{X}-p(q(\mathbf{X};\theta_e);\theta_d)||_2^2 \tag{2.1}$$

For the rest of the thesis, we consider that both the encoder and decoder are symmetric, meaning that they have the same number of layers, L, and units in each layer. We denote by $(q_l)_{l\in[\![1,L]\!]}$ the set of encoding layers such that $q_1 \circ q_2 \circ ... \circ q_L = q$ and by $(p_l)_{l\in[\![1,L]\!]}$ the set of encoding layers such that $p_1 \circ p_2 \circ ... \circ p_L = p$ [93].

A standard autoencoder can be enhanced by introducing an additional constraint, as in [248]. The idea is to add a reconstruction constraint on the intermediate layers.

$$\mathcal{L}_{recon} = \sum_{l=1}^{L} \alpha_l ||q_l(\mathbf{X};\theta_e) - p_{L+1-l}(q(\mathbf{X};\theta_e);\theta_d)|| \tag{2.2}$$

Where α_l is the weight of the constraint at layer l.

However, it is essential to avoid constraining on too many layers in order to avoid overfitting. We will further discuss this when presenting the design protocol.

Variational Autoencoders

Another example is Variational Autoencoders (VAEs). It is a deep generative model that can learn meaningful data representation from high-dimensional input data. This study will see this architecture as an extension of standard autoencoders in which the encoder encodes the input as a distribution over a latent space instead of a single point.

In this case, the encoding function q represents a variational distribution (known as encoding distribution) $q_\phi(\mathbf{Z}|\mathbf{X})$ in which ϕ is the learning parameter (denoted θ_e previously) and the decoding function p represents the posterior $p_\theta(\mathbf{X}|\mathbf{Z})$.

As told before, the particularity of a VAE is the ability to encode a distribution, which is represented by the fact that after the encoding phase, there is a sampling phase in which we sample points from the distribution $q_\phi(\mathbf{Z}|\mathbf{X})$ [136, 201].

Traditionally, the distributions in the VAE architecture are estimated as Gaussian: the encoder function will learn the two parameters μ and σ (respectively the mean and covariance) of the distribution $q_\phi(\mathbf{Z}|\mathbf{X})$ and will then reconstruct the input matrix using a reparametrization trick $z = \mu + \sigma\epsilon$

where $\epsilon \sim \mathcal{N}(0, I)$.

The loss function for this architecture can be written as the sum of two distinct losses. First, a reconstruction loss that focuses on the autoencoder's ability to reconstruct the data:

$$\mathcal{L}_{recon} = \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{X})}[p_\theta(\mathbf{X}|\mathbf{Z})] \tag{2.3}$$

This loss can be interpreted as the conditional entropy of $\mathbf{X}$ over $\mathbf{Z}$, which quantifies the amount of uncertainty that one has over the joint distribution $\mathbf{X}, \mathbf{Z}$ knowing $\mathbf{Z}$. More practically, it is related to the quantity of information of $\mathbf{X}$ that we can obtain from $\mathbf{Z}$ (qualifying a reconstruction potential).

Second, there is a regularization loss that aims at getting the encoding distribution as close as possible to the actual distribution of the latent vector. Traditionally, a Kullback-Leibler divergence is used:

$$\mathcal{L}_{reg} = D_{KL}(q_\phi(\mathbf{Z}|\mathbf{X})||p_\theta(\mathbf{Z})) \tag{2.4}$$

The total loss will benefit from the framework introduced in [115], which added weight to the regularization term to balance the two parts of the loss function depending on what we want to achieve.

$$\mathcal{L} = \mathcal{L}_{recon} + \beta\mathcal{L}_{reg} \tag{2.5}$$

2.1.3 . Multimodal Representation Learning

Representation learning has become a cornerstone in multimodality, enabling the integration of diverse data types into cohesive models that capture complex, cross-modal relationships. Representation learning is pivotal in the context of multimodal data because it transforms raw data from different modalities—such as images, text, and various omics data—into meaningful, low-dimensional representations that can be effectively combined and analyzed. This process facilitates a more comprehensive understanding of complex biological systems and improves the predictive power of models in biomedical research.

Autoencoders, particularly in their multimodal forms, are at the forefront of integrating different data types into a unified latent space. These models have been instrumental in learning joint representations from heterogeneous data sources, capturing the shared information across modalities while preserving unique characteristics. For instance, Multimodal Deep Denoising Autoencoders

(MDDA) have been effectively employed to integrate genomic, transcriptomic, and imaging data, thereby enhancing the accuracy of disease classification and patient outcome prediction [184, 17].

Deep generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have recently gained significant success in the unsupervised learning of latent representations from high-dimensional and structured data like images, audio, and text [94, 136, 201]. These models are crucial for data synthesis and play a fundamental role in data analysis and transformation. An effective learned representation must capture high-level characteristics that remain invariant to small, local changes in the input data and should be as disentangled as possible for enhanced explainability. Hierarchical and disentangled generative models have shown their effectiveness in addressing downstream learning tasks, demonstrating the potential to improve predictive accuracy and model interpretability [23, 244]. Furthermore, introducing more advanced models, such as MDVAE proposed in [209], which leverages hierarchical disentanglement in generative processes, has further pushed the boundaries of generative models in unsupervised learning.

Convolutional Neural Networks (CNNs) have significantly contributed to multimodal representation learning, mainly when one of the modalities is image-based. CNNs are adept at extracting high-level features from visual data, which can be combined with other modalities, such as text, to create comprehensive joint representations. These combined representations have led to advances in cancer subtype classification, where integrating histopathology images with gene expression data has significantly improved diagnostic accuracy [144, 281].

Beyond individual model types, specialized multimodal architectures have been developed to address the challenges of integrating diverse data sources. These architectures typically involve parallel processing networks for each modality, followed by a fusion layer that combines the representations into a unified model. Techniques such as Deep Canonical Correlation Analysis (DCCA) [9], Multimodal Factorization Models (MFM) [240], and Regularized Generalized Canonical Correlation Analysis (RGCCA) [92] have shown promise in capturing the intricate correlations between modalities, thus enhancing the overall predictive performance of multimodal systems. RGCCA, in particular, extends the classical CCA by incorporating regularization and generalization components, making it highly effective in scenarios where data is complex and possibly incomplete.

Despite the significant advancements, challenges remain in multimodal representation learning, particularly in effectively aligning disparate modalities and managing incomplete data. Furthermore,

the interpretability of these complex models is crucial, especially in clinical applications where understanding the rationale behind predictions is vital. Addressing these challenges will require the development of more interpretable models and novel strategies for seamless data integration, setting the stage for the next generation of multimodal learning techniques.

2.2 . Multimodal Integration Strategies

Combining different data types or modalities is fundamental in various machine learning applications, as integrating information from sources such as text, images, and audio can significantly improve model accuracy and robustness. This multimodal integration is crucial in fields like multimedia processing, robotics, natural language processing, and medicine, where complex and diverse data sources are shared. The primary strategies for merging these diverse data types include early, late, and joint integration, each offering distinct advantages and facing unique challenges depending on the specific context and objectives. Importantly, all three multimodal integration approaches can be applied effectively to predictive modeling and representation learning tasks.

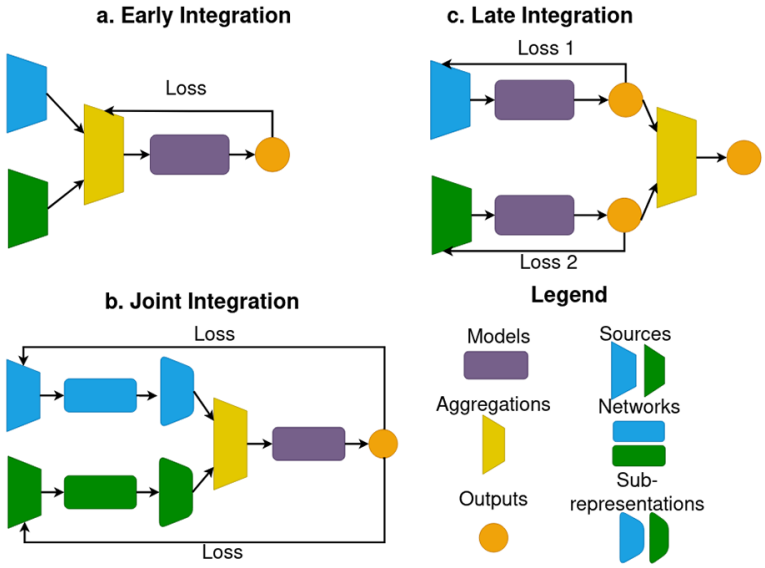


Figure 2.1: *An overview of the different integration strategies*

2.2.1 . Early Integration (Feature-Level Integration)

Early or feature-level integration merges various data types prior to analysis or modeling. This approach combines various datasets into one thorough feature space right at the start of the processing pipeline.

Suppose we have datasets $X_1, X_2, \ldots, X_n$ corresponding to different modalities. The integrated dataset X can be represented as:

$$X = [X_1 \quad X_2 \quad \ldots \quad X_n]$$

where $[\cdot]$ denotes the concatenation of features across different modalities [18].

This method enables models to learn the relationships between features from various modalities, resulting in a more detailed and insightful representation. However, the main problem is the large number of dimensions, which may lead to overfitting and higher computational costs. Moreover, compatible preprocessing is required for concatenating all modalities [199].

For representation learning with autoencoders and VAEs, for example, data from all modalities are concatenated and input into the network:

$$X = [X_1 \quad X_2 \quad \ldots \quad X_n]$$

The encoder transforms X into a latent space Z:

$$Z = \text{Encoder}(X)$$

The decoder then attempts to reconstruct X from Z:

$$\hat{X} = \text{Decoder}(Z)$$

The loss function typically used is the mean squared error (MSE) for autoencoders or the evidence lower bound (ELBO) for VAEs (more details about its computation are found in Appendix A.4), incorporating both reconstruction loss and KL divergence. Early integration for autoencoder architectures is exemplified through the SDAE architecture. In the rest of the study, we will denote by EI-AE and EI-VAE, respectively, the early integration autoencoder and variational autoencoder.

2.2.2 . Late Integration (Decision-Level Integration)

Late integration, or decision-level integration, entails training distinct models for individual data types and merging their outputs or decisions during a subsequent phase. Each model extracts information separately from its type, and the final result combines all outputs.

Let $f_1, f_2, \ldots, f_n$ be the models trained on datasets $X_1, X_2, \ldots, X_n$ respectively. The final output y can be derived from:

$$y = \text{Combine}(f_1(X_1), f_2(X_2), \ldots, f_n(X_n))$$

Where Combine could be a function such as averaging, a weighted sum, or a more complex decision function like a neural network or a voting scheme [18, 210].

This method reduces the chance of overfitting by handling each data type separately, enabling freedom in choosing and optimizing models for each data type. However, it might not be able to encompass crucial connections between various modes necessary for complete comprehension or accurate forecasting [199].

For representation integration, we can imagine separate autoencoders/VAEs trained on each modality:

$$Z_i = \text{Encoder}_i(X_i), \quad \hat{X}_i = \text{Decoder}_i(Z_i)$$

The encoded representations $Z_1, Z_2, \ldots, Z_n$ are combined using a model g that predicts the final output Y:

$$Y = g(Z_1, Z_2, \ldots, Z_n)$$

The model g could be trained to optimize a specific task-related performance metric, integrating insights from each modality. Late integration is exemplified for autoencoder architectures through a hierarchical approach, as introduced by [225] and represented in Fig. 2.2.

2.2.3 . Joint Integration (Model-Level Integration)

Joint integration, also known as model-level integration, refers to models created to consider interactions between various data types during the learning process. This approach uses designs integrating multi-task learning or common layers to deal with several modalities simultaneously.

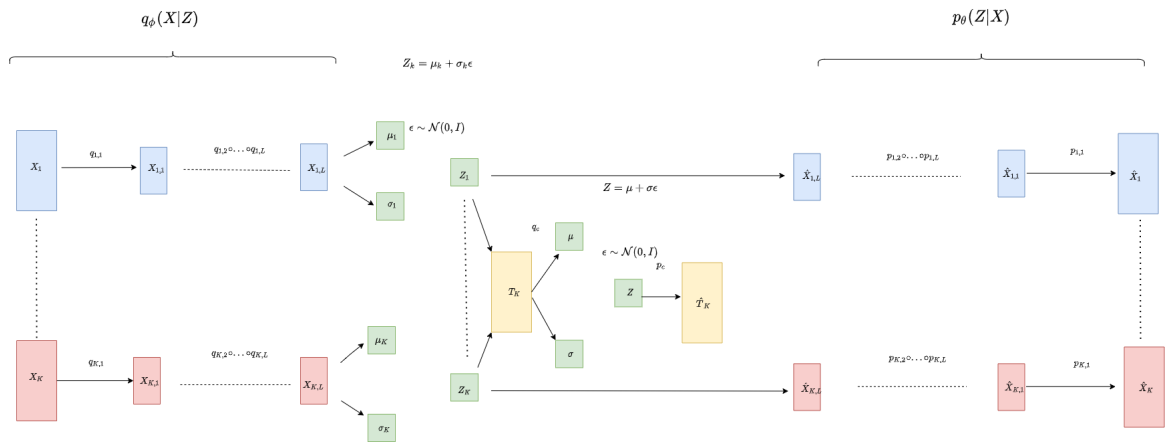


Figure 2.2: ***H-VAE Architecture :*** *This figure depicts the Hierarchical Variational Autoencoder (H-VAE), which utilizes a late integration approach for multimodal representation learning. Each modality $X_1, \ldots, X_K$ is first encoded into separate latent variables $Z_1, \ldots, Z_K$. These are then combined into a shared latent variable Z, capturing the joint distribution across all modalities. The shared latent representation is used to reconstruct each modality through respective decoders.*

Consider a model with shared parameters θ and modality-specific parameters $\theta_1, \theta_2, \ldots, \theta_n$. The model can be formulated as:

$$y = g(X_1, X_2, \ldots, X_n; \theta, \theta_1, \theta_2, \ldots, \theta_n)$$

Where g is a complex architecture that processes and integrates information from all modalities, either sequentially or in parallel [184].

This method can effectively learn complex connections between different aspects, leading to better results when these connections are precious. However, this method necessitates complex model structures and usually larger datasets to be trained successfully without falling into overfitting [276].

For representation learning, a multi-branch architecture is used where each branch processes inputs from one modality and feeds into a shared encoder:

$$Z_i = \text{BranchEncoder}_i(X_i)$$

The shared encoder learns a joint representation:

$$Z = \text{Merge}(Z_1, Z_2, \ldots, Z_n)$$

The decoder reconstructs combined or separate modal outputs:

$$\hat{X} = \text{Decoder}(Z)$$

The merging function Merge could be as simple as concatenation or involve more complex interactions such as feature fusion through addition or multiplication layers. Examples of joint integration for autoencoder architectures include the Disjointed Deep Autoencoder (DDAE) [248] and the X-VAE architecture [225].

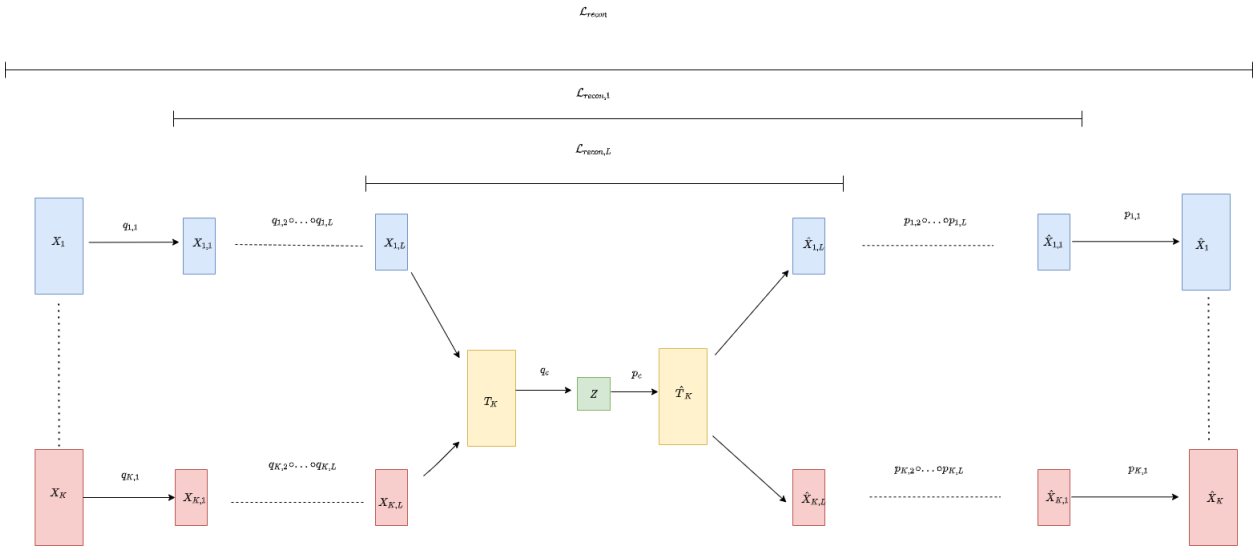


Figure 2.3: ***DDAE Architecture :*** *The figure illustrates the DDAE architecture, which employs a joint integration strategy for multimodal representation learning. In this model, data from different modalities $X_1, \ldots, X_K$ are encoded through respective layers into a shared latent space Z, facilitating the integration of multimodal information. Additionally, the architecture incorporates intermediate reconstruction losses $\mathcal{L}_{recon,1}, \ldots, \mathcal{L}_{recon,L}$ at various stages of the encoding process, ensuring that meaningful representations are learned at each layer. These losses help maintain the integrity of the information as it passes through the network, ultimately improving the quality of the reconstructed outputs $\hat{X}_1, \ldots, \hat{X}_K$.*

The choice of a multimodal integration approach depends on the specific task, data characteristics, and application goals. Each method has its strengths and limitations, and selecting the right one can significantly influence the effectiveness and efficiency of the learning process. Our objective is to identify the most effective strategies for integrating and interpreting data across different modalities, particularly in the context of representation learning. This analysis will guide the selection of optimal integration techniques for specific use cases. To do this, we will evaluate each integration strategy in

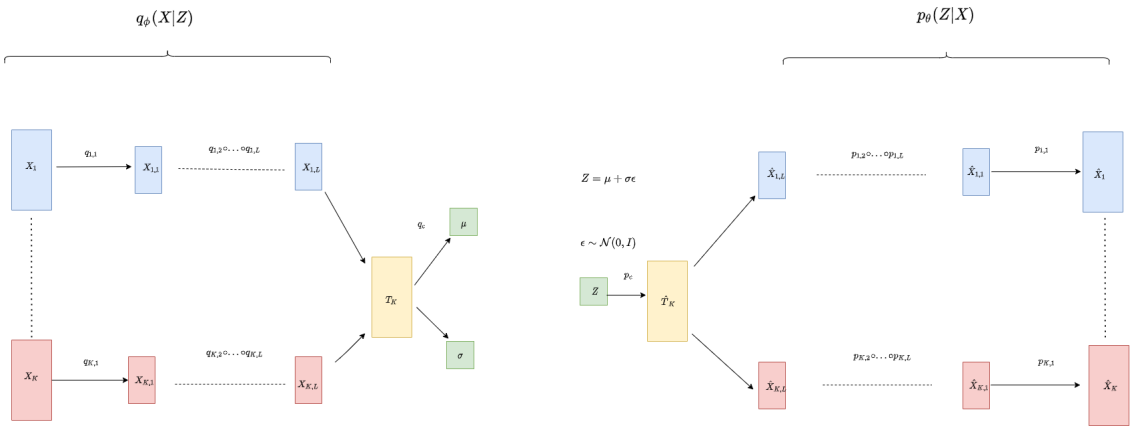


Figure 2.4: ***X-VAE Architecture :*** *The figure illustrates the architecture of a Variational Autoencoder (VAE) designed for multimodal representation learning using a joint integration strategy. In this model, data from different modalities $X_1, \ldots, X_K$ are encoded into respective latent representations $Z_1, \ldots, Z_K$ through modality-specific encoders. These latent representations are then combined in a shared latent space Z, characterized by the mean μ and standard deviation σ, which captures the joint distribution of all modalities. The shared latent variable Z is used to reconstruct each modality $\hat{X}_1, \ldots, \hat{X}_K$ via their respective decoders.*

the context of multi-omics integration.

2.3 . Application to Multi-Omics Integration

We aim to systematically assess and contrast various multimodal integration approaches, such as early, late, and joint integration. This involves utilizing two representation learning models: factorial techniques like Principal Component Analysis (PCA) and deep learning techniques such as autoencoders and variational autoencoders (VAEs) in the context of multi-omics integration.

2.3.1 . Related Work on Multi-Omics Integartion

Multi-omics integration has become a critical research focus in health sciences and precision medicine, offering new insights into complex biological processes, particularly in cancer research. The integration of diverse omics data types is essential for understanding the intricate molecular mechanisms underlying diseases. Various statistical and deep learning methods have been developed to address the challenges of multi-omics integration, each offering unique advantages and facing distinct challenges.

One of the foundational approaches in this field is Principal Component Analysis (PCA), exten-

sively explored in various studies, including its adaptations such as Multiple Factor Analysis (MFA), Consensus PCA, and multi-block PCA [127, 215, 112]. MFA is particularly suited for integrating multiple data tables, each representing a different modality. It extends PCA by simultaneously analyzing several datasets, preserving the structure of each modality while capturing common patterns across them. MFA can be applied in both early and joint integration strategies, depending on the desired outcome. Kernel Multiple Factor Analysis (KFA) further enhances MFA by incorporating kernel methods, enabling the analysis of nonlinear relationships between modalities [1, 188, 223, 82, 278].

Another significant method is Non-negative Matrix Factorization (NMF), which, unlike PCA, imposes a non-negative constraint rather than an orthogonality one. NMF has been widely used for multi-omics data integration, particularly in cancer research, where it aids in uncovering meaningful biological patterns [272]. Additionally, Joint Dimensionality Reduction (jDR) methods have emerged as a valuable extension of traditional factorial methods, with applications in clustering and module identification in disease-associated mechanisms [39, 245, 32]. These methods include Bayesian approaches that rely on assumptions about data distributions and network-based methods that use graph representations to identify modules within complex biological systems.

Deep learning has increasingly been applied to multi-omics integration, driven by its success in other medical applications such as imaging and diagnostics. Autoencoders, for instance, have been extensively used for multi-omics data integration. Chaudhary et al. utilized autoencoders for survival prediction, while other approaches like OmiVAE and OmiEmbed have improved representation and multitask learning, respectively [207, 44, 274, 273]. Simidjievski et al. explored variational frameworks for various tasks, highlighting the flexibility of deep learning architectures in handling multi-omics data [225]. Other notable models include the Salmon framework, which integrates multi-omics data using neural networks for survival analysis [118], and the OmiVAE framework, which was specifically applied to study ovarian cancer [117].

Recent advancements have introduced sophisticated models like MOGONET (Multi-Omics Graph Convolutional Network), which uses graph convolutional networks for supervised classification tasks, demonstrating significant promise in biomedical classification by integrating multiple omics types [253]. Another notable development is the Knowledge Distillation and Supervised Variational Autoencoders (KD-SVAE-VCDN), which combines knowledge distillation with supervised variational autoencoders to handle incomplete multi-omics data, making it particularly useful for disease progression

prediction [255]. Additionally, the Weighted Affinity and Self-Diffusion Network Integration method has been developed to enhance network connections and improve clustering performance for cancer subtype identification [155].

2.3.2 . Experimental Setup

Classification on Ovarian Subtypes

Ovarian cancer is one of the most common gynecological cancers with the highest mortality rate due to the absence of early-stage symptoms. It is molecularly heterogeneous and can be classified into four molecular subtypes: Mesenchymal (C1), Immunoreactive (C2), Differentiated (C4), and Proliferated (C5). A better understanding of these subtypes is essential for improved prognosis and therapy.

We used the TCGA-OV cohort to evaluate the classification performance of our models by classifying the samples according to the ovarian molecular subtypes. The dataset includes CNV, miRNA, and DNA methylation data. More details of TCGA are present in appendix B.1.

Table 2.1: ***Classification on Ovarian Subtypes: Data Description***

Omic Type	Omic Data	Feature Size	Sample Size
Genomics	CNV	24,776	579
Transcriptomics	mRNA	12,042	593
Epigenomics	DNA methylation	21,666	616

For molecular subtypes, we used the ConsensusOV R package and a neural network classifier optimized for each design.

Survival Analysis for Breast Cancer

We used the ovarian cancer TCGA cohort, TCGA-OV, for this test case. We utilized omics data and clinical annotations to perform survival prediction using the DeepSurv model, a non-linear approach to the Cox Proportional Hazard model.

The classical hazard function is defined as follows:

$$\mu(t, x_i) = \mu_0(t)\Psi(x_i)$$

Where $\Psi(x_i) = \exp(\psi(x_i))$, with ψ being a nonlinear risk function. The model uses Efron's negative log-likelihood formula:

$$L(\theta) = -\sum_{i:E_i=1} (\hat{\mu}(t, x_i; \theta) - \log \sum_{j \in \mathcal{R}(T_i)} e^{\hat{\mu}(x_i;\theta)})$$

Where $\hat{\mu}(x; \theta)$ is the risk function estimated by the output layer of the network. $\mathcal{R}(t)$ is the risk set, the set of patients still at risk of failure after time t. More details on Survival Analysis can be found in Appendix A.3

Evaluation Strategy

To rigorously assess the performance of the different models, we employ a 5-fold cross-validation strategy. This involves partitioning the dataset into five equal-sized folds, ensuring each fold is used as a test set once while the remaining four folds serve as the training set. This process is repeated five times, enabling every data point to be included in both the training and test phases.

As there are missing samples in each modality, we only consider samples that have all the modalities for the evaluation.

For each fold, we compute several evaluation metrics to ensure a comprehensive analysis of model performance. The metrics considered include balanced accuracy, which provides a balanced measure of accuracy accounting for imbalanced class distributions, and the area under the curve (AUC), which evaluates the model's ability to distinguish between classes.

For survival analysis, we use the concordance index (C-index), measuring the model's discriminative power by quantifying the concordance between the predicted and actual event times, and the integrated Brier score (IBS), which assesses the accuracy of probabilistic predictions by measuring the mean squared differences between the observed survival outcomes and the predicted probabilities over time (Details of those metrics can be found in Appendix A.3). After performing 5-fold cross-validation, we aggregate the results by calculating the mean and standard deviation of each metric across all folds. This approach provides a robust estimate of model performance, accounting for variability and ensuring the reliability of our evaluation, thereby facilitating informed improvements and refinements.

2.3.3 . Results

The evaluation of multimodal integration strategies for classifying ovarian subtypes and predicting survival outcomes in the TCGA-OV cohort, shown in tables 2.2 and 2.3 reveals a clear hierarchy of

Table 2.2: ***Classification of Ovarian subtypes*** *: Evaluation of multimodal integration startegies using multiple models evaluated on the classification of ovarian subtypes in the TCGA-OV cohort. The evaluation is done using balanced accuracy along with the Area under ROC curve (AUC)*

Model	Accuracy	AUC
MFA	0.743 ± 0.053	0.941 ± 0.024
Kernel MFA	0.754 ± 0.041	0.942 ± 0.022
EI-AE	0.793 ± 0.038	0.953 ± 0.018
EI-VAE	0.822 ± 0.029	0.963 ± 0.016
DDAE	0.844 ± 0.034	0.971 ± 0.012
X-VAE	$\mathbf{0.872 \pm 0.021}$	$\mathbf{0.981 \pm 0.014}$
H-AE	0.613 ± 0.081	0.865 ± 0.032
H-VAE	0.844 ± 0.033	0.962 ± 0.013

Table 2.3: ***Survival Analysis on Ovarian cancer*** *: Evaluation of multimodal integration startegies using multiple models evaluated on the survival outcome prediction in the TCGA-OV cohort. The evaluation is done using concordance index (C-index) along with the Integrated Brier Score (IBS).*

Model	C-Index	IBS
MFA	0.535 ± 0.064	0.243 ± 0.046
Kernel MFA	0.541 ± 0.056	0.213 ± 0.041
EI-AE	0.601 ± 0.048	0.192 ± 0.032
EI-VAE	0.631 ± 0.039	0.176 ± 0.027
DDAE	0.651 ± 0.034	0.162 ± 0.024
X-VAE	$\mathbf{0.672 \pm 0.044}$	$\mathbf{0.143 \pm 0.034}$
H-AE	0.553 ± 0.085	0.242 ± 0.036
H-VAE	0.622 ± 0.041	0.181 ± 0.022

model performance, with joint integration models, particularly those employing variational autoencoders (VAEs), emerging as the most effective. Among these, the X-VAE model exhibits the highest accuracy and AUC for classification and the best C-index and lowest IBS for survival analysis, showcasing its superior capability in handling complex multimodal data.

Early integration models, such as EI-AE and EI-VAE, demonstrate moderate performance, with EI-VAE outperforming EI-AE. This suggests that while early integration captures some inter-modality relationships by merging data at the input level, it may need to exploit the rich interactions present across modalities fully. VAEs enhance this by providing a more sophisticated representation of the integrated data, leading to better outcomes than standard autoencoders.

Late integration models, like H-AE and H-VAE, generally show poorer performance, with H-AE being the least effective across all metrics. H-VAE, however, shows a marked improvement over H-AE,

underscoring the advantages of using VAEs. The main shortcoming of late integration is its failure to capture inter-modality interactions effectively. By processing each modality separately before combining their outputs, late integration strategies often miss the synergistic information that could be leveraged if the modalities were integrated earlier. This separate processing can result in losing complementary information crucial for accurate predictions.

The good performance of joint integration models, such as DDAE and X-VAE, can be attributed to their ability to integrate data modalities simultaneously. This approach allows for extracting comprehensive and correlated features, significantly enhancing predictive accuracy. Joint integration models can dynamically capture interactions between different data modalities during learning, leading to a more holistic understanding of the data. VAEs, in particular, are adept at learning complex data distributions and generating informative latent spaces, further boosting their performance in classification and survival prediction.

Additionally, the ability of joint integration models to leverage inter-modality relationships more effectively than early and late integration methods could explain their superior performance. In early integration, the data modalities are merged at the input level, which can obscure the distinct contributions of each modality and lead to suboptimal feature extraction. Conversely, late integration processes each modality independently, failing to capture the synergistic interactions that occur when modalities are considered together from the outset. This independent processing can diminish the potential predictive power derived from the complementary nature of multimodal data.

Moreover, the distinct advantage of variational autoencoders over traditional autoencoders is evident across all integration strategies. VAEs can learn richer, more nuanced data representations, which translates to improved performance in classification and survival analysis tasks. Their ability to model complex distributions and generate informative latent spaces allows for better capture of the underlying data structure, leading to more accurate and robust predictions. This highlights the critical role of advanced modeling techniques in enhancing the efficacy of multimodal integration strategies.

To further test the limits of each integration strategy, an additional experiment was conducted using multiple combinations of omics data. The results displayed in Tables 2.4 and 2.5 reveal that late integration models, like H-VAE, achieved their best performance with single omic data, particularly mRNA, rather than with combinations of multiple omics. This suggests that late integration strategies are not well-suited to capturing cross-modality signals, highlighting a significant limitation in their

Table 2.4: ***Classification of Ovarian Subtypes**: Evaluation of different integration strategies using VAEs on multiple combinations of omics data.*

Model	Data Combination	Accuracy	AUC
EI-VAE			
	CNV	0.713 ± 0.043	0.921 ± 0.022
	mRNA	0.819 ± 0.029	0.961 ± 0.018
	Methyl	0.753 ± 0.034	0.941 ± 0.021
	CNV + mRNA	0.809 ± 0.031	0.931 ± 0.019
	CNV + Methyl	0.772 ± 0.038	0.943 ± 0.020
	mRNA + Methyl	$\mathbf{0.822 \pm 0.029}$	$\mathbf{0.963 \pm 0.016}$
	CNV + mRNA + Methyl	0.821 ± 0.033	0.953 ± 0.018
X-VAE			
	CNV	0.741 ± 0.032	0.931 ± 0.020
	mRNA	0.871 ± 0.021	0.979 ± 0.013
	Methyl	0.782 ± 0.029	0.951 ± 0.018
	CNV + mRNA	0.864 ± 0.023	0.972 ± 0.015
	CNV + Methyl	0.792 ± 0.031	0.952 ± 0.017
	mRNA + Methyl	$\mathbf{0.872 \pm 0.021}$	$\mathbf{0.981 \pm 0.014}$
	CNV + mRNA + Methyl	0.861 ± 0.022	0.973 ± 0.015
H-VAE			
	CNV	0.641 ± 0.048	0.891 ± 0.025
	mRNA	$\mathbf{0.844 \pm 0.033}$	$\mathbf{0.962 \pm 0.013}$
	Methyl	0.693 ± 0.044	0.911 ± 0.023
	CNV + mRNA	0.802 ± 0.037	0.943 ± 0.021
	CNV + Methyl	0.722 ± 0.041	0.913 ± 0.022
	mRNA + Methyl	0.813 ± 0.032	0.943 ± 0.019
	CNV + mRNA + Methyl	0.813 ± 0.035	0.944 ± 0.021

design.

Moreover, joint integration models like X-VAE, while generally performing well, showed decreased performance when CNV data was included. This indicates that classical integration strategies may need help effectively integrating CNV data with other omics data. The reduced efficacy of CNV integration suggests that these strategies may need to be fully equipped to simultaneously handle the complexities and nuances of all types of omics data.

Additionally, as illustrated in Figure 2.5, the loss function for RNAseq data exhibits a more rapid convergence than that for CNV data. This disparity in convergence rates impacts the multimodal loss, which appears to align more closely with the RNAseq convergence trajectory. This observation suggests that the CNV data may not be adequately integrated within the multimodal framework, po-

Table 2.5: ***Survival Analysis on Ovarian Cancer**: Evaluation of different integration strategies using VAEs on multiple combinations of omic data*

Model	Data Combination	C-Index	IBS
EI-VAE			
	CNV	0.571 ± 0.053	0.233 ± 0.041
	mRNA	$\mathbf{0.631 \pm 0.039}$	$\mathbf{0.176 \pm 0.027}$
	Methyl	0.601 ± 0.048	0.192 ± 0.032
	CNV + mRNA	0.612 ± 0.042	0.182 ± 0.029
	CNV + Methyl	0.591 ± 0.047	0.194 ± 0.031
	mRNA + Methyl	0.621 ± 0.037	0.176 ± 0.026
	CNV + mRNA + Methyl	0.621 ± 0.041	0.181 ± 0.028
X-VAE			
	CNV	0.611 ± 0.041	0.203 ± 0.032
	mRNA	0.662 ± 0.044	0.143 ± 0.034
	Methyl	0.621 ± 0.038	0.183 ± 0.031
	CNV + mRNA	0.662 ± 0.043	0.153 ± 0.033
	CNV + Methyl	0.631 ± 0.042	0.184 ± 0.030
	mRNA + Methyl	$\mathbf{0.672 \pm 0.044}$	$\mathbf{0.143 \pm 0.034}$
	CNV + mRNA + Methyl	0.661 ± 0.040	0.154 ± 0.032
H-VAE			
	CNV	0.551 ± 0.057	0.243 ± 0.041
	mRNA	0.642 ± 0.041	0.181 ± 0.022
	Methyl	0.572 ± 0.052	0.212 ± 0.036
	CNV + mRNA	0.612 ± 0.046	0.191 ± 0.033
	CNV + Methyl	0.581 ± 0.049	0.213 ± 0.035
	mRNA + Methyl	0.632 ± 0.042	0.182 ± 0.031
	CNV + mRNA + Methyl	0.611 ± 0.045	0.192 ± 0.033

tentially due to insufficient convergence time within the joint integration strategy.

This observation raises concerns about the robustness and adaptability of current integration methods. Suppose joint integration, otherwise the most effective strategy, fails to improve or maintain performance by adding CNV data. In that case, it underscores the need for developing more sophisticated or tailored integration approaches that can better accommodate the diverse nature of omics data. The results clearly show that while variational autoencoders and joint integration offer substantial benefits, they also have limitations that must be addressed to achieve genuinely comprehensive and effective multimodal data integration.

2.3.4 . Conclusion

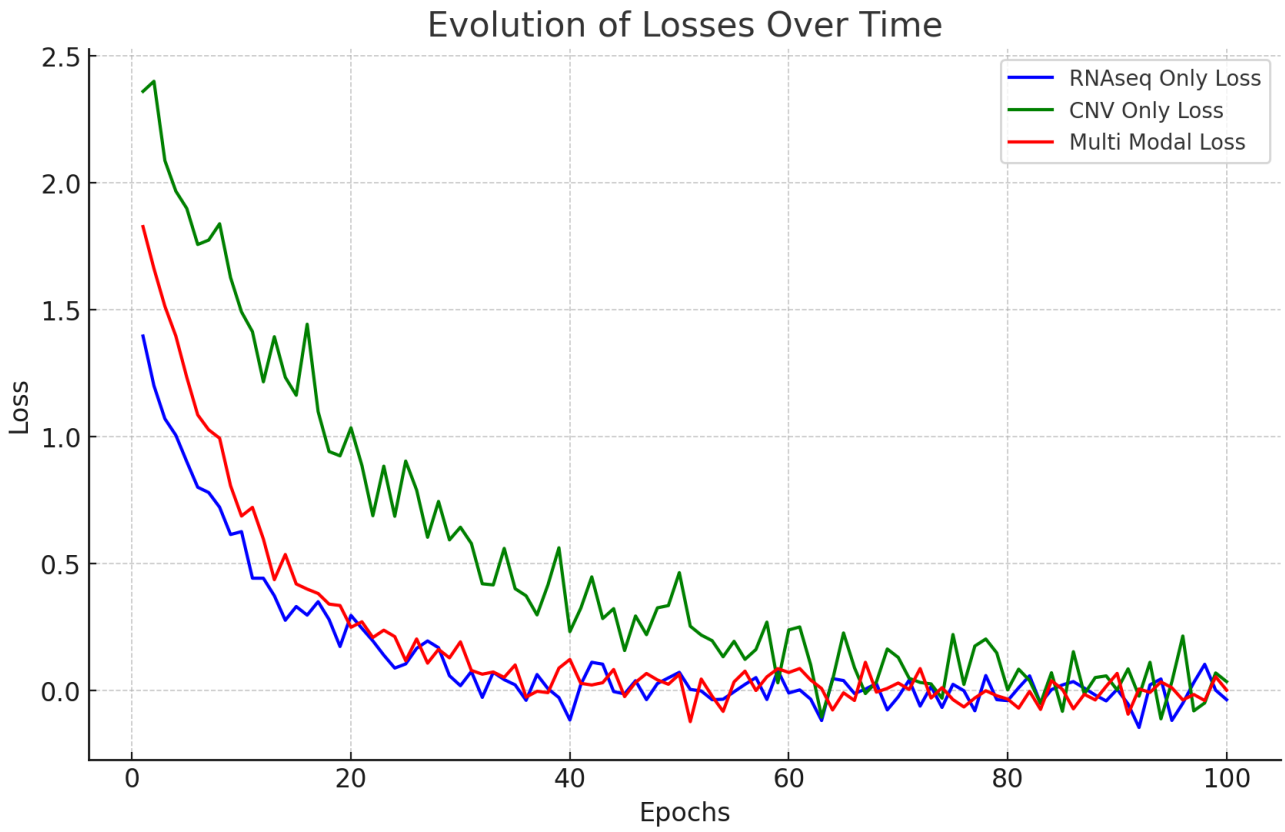


Figure 2.5: *Comparison of the loss evolution between CNV, RNAseq and Multimodal model for Joint Integration for the TCGA-OV classification task*

This chapter highlighted the superior performance of variational autoencoders (VAEs) over traditional factorial methods and autoencoders for multimodal data integration, particularly in classifying ovarian subtypes and predicting survival outcomes in the TCGA-OV cohort. Joint integration models, like X-VAE, consistently outperformed early and late integration approaches, especially with mRNA and methylation data. However, challenges were noted with late integration strategies and the effective incorporation of CNV data, indicating limitations in current methodologies. These findings underscore the need for more sophisticated integration techniques that can handle the complexity of diverse omics data. The next chapter will expand this investigation to include a broader range of datasets and tasks, aiming to test the robustness of these integration strategies and develop improved methods to overcome the limitations identified.

Multi-Omics Integration for Precision Medicine

Contents

Abstract

Integrating multi-omics data presents significant challenges due to the complexity and volume of datasets across genomics, transcriptomics, proteomics, and epigenomics. To address these challenges, we introduced CustOmics, a deep learning framework designed to efficiently and effectively integrate multi-omics data and validate it using multiple TCGA data. This chapter also details the application of CustOmics to Myelodysplastic Syndromes (MDS), which has enabled the identification of novel biomarkers and potential therapeutic targets that were not detectable through traditional methods. Our findings demonstrate the potential of CustOmics to enhance the accuracy of multi-omics data interpretation and advance precision medicine in oncology.

Multi-omics integration in oncology seeks to combine various data types—such as genomics, transcriptomics, proteomics, and metabolomics—to enhance disease understanding and develop personalized treatment plans. This approach holds great promise for advancing personalized medicine by identifying novel biomarkers, unraveling complex disease pathways, and tailoring treatments to individual patients. However, as demonstrated in the previous chapter, classical multimodal integration techniques—including early, joint, and late integration—fail to optimally integrate multi-omics data, highlighting significant limitations in current methodologies.

These challenges are substantial: the vast amount of data and its intricate nature demand advanced computational resources and sophisticated analytical methods for effective handling and processing. Additionally, the diversity in data types, scales, and biological variability complicates standardization efforts, making comparisons and integrations difficult. Widespread missing data, often resulting from varying experimental techniques and detection limits, further complicates analyses, necessitating reliable methods to address these gaps without introducing bias. Consequently, there is a pressing need for novel integration techniques to overcome these challenges and advance precision oncology by ensuring the biological interpretability of integrated data.

Numerous research initiatives have generated valuable data from various molecular sources. For example, The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) has analyzed numerous tumor samples using diverse molecular assays, providing data such as genome sequencing, RNA sequencing, DNA methylation, and proteomics. Integrating this diverse data is crucial to minimize the uncertainty of different experimental conditions and uncover interactions that a single data source cannot reveal.

This broad spectrum of data presents two significant challenges. The first relates to the data's high dimensionality. Due to the complex genetic makeup of human molecular profiles, omics data often suffers from the 'curse of dimensionality,' where the number of features exceeds the number of samples [174]. In such high-dimensional spaces, interrelated features can lead to significant overlap, reducing the predictive accuracy of algorithms. The second challenge is data heterogeneity, which arises from various sources and represents different aspects of biological systems in human omics data [27]. For instance, transcriptomics and proteomics data are normalized and scaled differently, resulting in varying data ranges and distributions; metabolomics data can also exhibit sparsity due to variables falling below detection limits and being recorded as null values.

To address these challenges and mitigate overfitting in predictive tasks, conventional approaches often involve handpicking a limited number of molecular features based on domain expertise [37] or applying dimensionality reduction methods before further analysis. Other methods also incorporate penalties when building the model [25]. However, these methods risk missing critical underlying patterns across the genome due to the significant disparities between data sources. This highlights the urgent need for more flexible and advanced integration methods. Building on the observations from the previous chapter, this chapter focuses on the conception, evaluation, and application of a novel integration technique designed to overcome the shortcomings of classical methods, aiming to improve the robustness and effectiveness of multi-omics data integration in precision oncology.

3.1 . CustOmics: A versatile deep-learning based strategy for multi-omics integration

The vast potential of multi-omics data in cancer research involves different omic data types to offer a thorough understanding of the molecular processes involved in cancer. However, existing methods for combining these various data formats frequently need to improve their capability to manage the intricacy and size of the data efficiently. This difference highlights the urgent requirement for better approaches to utilize the full scope of multi-omics data effectively.

As highlighted in Section 2.3, all existing multimodal integration techniques struggle to effectively integrate RNAseq data with other omics data because of the significant differences in predictive signals. To tackle this problem, we introduce one of this thesis's main contributions, CustOmics, an

innovative deep learning system explicitly created to combine multi-omics information. CustOmics seeks to address the constraints of current methods by utilizing an integration strategy tailored to leverage various types of data effectively.

3.1.1 . Method

To address the limitations inherent in joint and late integration strategies, we propose a new approach, "mixed integration." This strategy is designed to harness the strengths of both joint and late integration while mitigating their respective weaknesses. By combining elements of each, mixed integration aims to create a more robust and practical framework for data analysis, capitalizing on the synergies between the two while addressing their shortcomings.

This integration strategy will be the foundation for building a customizable architecture for multi-omics integration, CustOmics. The proposed method is a hierarchical mixed-integration consisting of an autoencoder for each source, creating a sub-representation that will then be fed to a central variational autoencoder. This new integration strategy benefits from two training phases. The first phase will act as a normalization process: each source will train separately to learn a more compact representation that synthesizes its information with less noise. This will help the integration as we will lose all imbalance issues between the sources and avoid losing focus when a source has an inferior dimensionality or weaker signal than the others. The second phase will constitute a simple joint integration between the learned sub-representations while still training all the encoders to fine-tune those representations, as some signals are enhanced in the presence of other sources.

Regarding the regularization loss for the central layer, the KL divergence can be an obstacle to generalization. As stated in [277], the KL divergence suffers from various problems. First, the model can fail to learn a meaningful input representation. Indeed, the KL divergence can sometimes be too restrictive and naturally tends to make the latent code a random sample from $p_\theta(\mathbf{z})$. The second is that the KL divergence can make the model overfit and learn a latent code with variance tending to infinity.

So, instead, we will use the Maximum Mean Discrepancy (MMD) to assess the distance between the distributions. This distance stands on the foundation that two distributions are identical if their

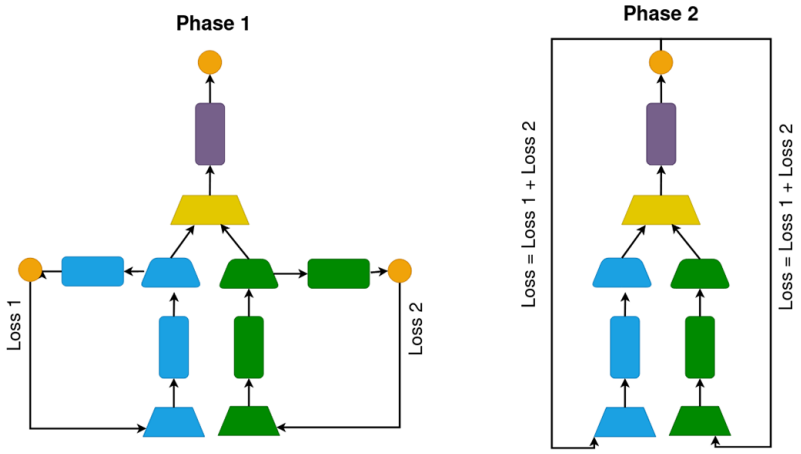


Figure 3.1: ***Mixed Integration :*** *This figure illustrates a mixed integration strategy for multimodal learning, introduced in two distinct phases. In Phase 1, each modality is trained separately, with the merging layer (yellow) kept frozen. This phase focuses on optimizing the individual modality-specific networks, resulting in two separate loss functions (Loss 1 and Loss 2) that are independently minimized. In Phase 2, the merging layer is unfrozen, allowing the modalities to be trained jointly. In this phase, the loss functions are combined into a single loss (Loss = Loss 1 + Loss 2), facilitating the integration of information across modalities and improving the overall model's performance by capturing inter-modality interactions.*

moments are the same. Let p, q be two distributions; the MMD distance is given as follows:

$$\begin{aligned} MMD(p(x)||q(x)) =& \mathbb{E}_{p(x),p(x')}(\kappa(x,x')) + \mathbb{E}_{q(x),q(x')}(\kappa(x,x')) \\ & - 2\mathbb{E}_{p(x),q(x')}(\kappa(x,x')) \end{aligned}$$

where κ is a kernel function and where x and x' are two sample points. We will choose a Gaussian kernel $\kappa(x,x') = e^{\frac{-||x-x'||^2}{2\sigma^2}}$.

Whereas all the models mentioned previously build the latent representation unsupervised, we also create latent features adapted to specific tasks like classification or survival. This idea has been used multiple times in the literature, for example, in [273]. The solution relies on adding a task-related loss to the autoencoder objective function. Therefore, we denote by $\mathcal{L}_{task}$ the loss such that the total loss function would be expressed as follows: $\mathcal{L}_{tot} = \mathcal{L}_{AE} + \alpha\mathcal{L}_{task}$, where $\mathcal{L}_{AE}$ is the autoencoder loss.

For the classification task, we use a categorical cross-entropy loss defined by $\mathcal{L}_{class} = \sum_i y_i log(\hat{y}_i)$, where y_i is the ground truth for the i^{th} sample, and $\hat{y}_i$ its estimation with the downstream model.

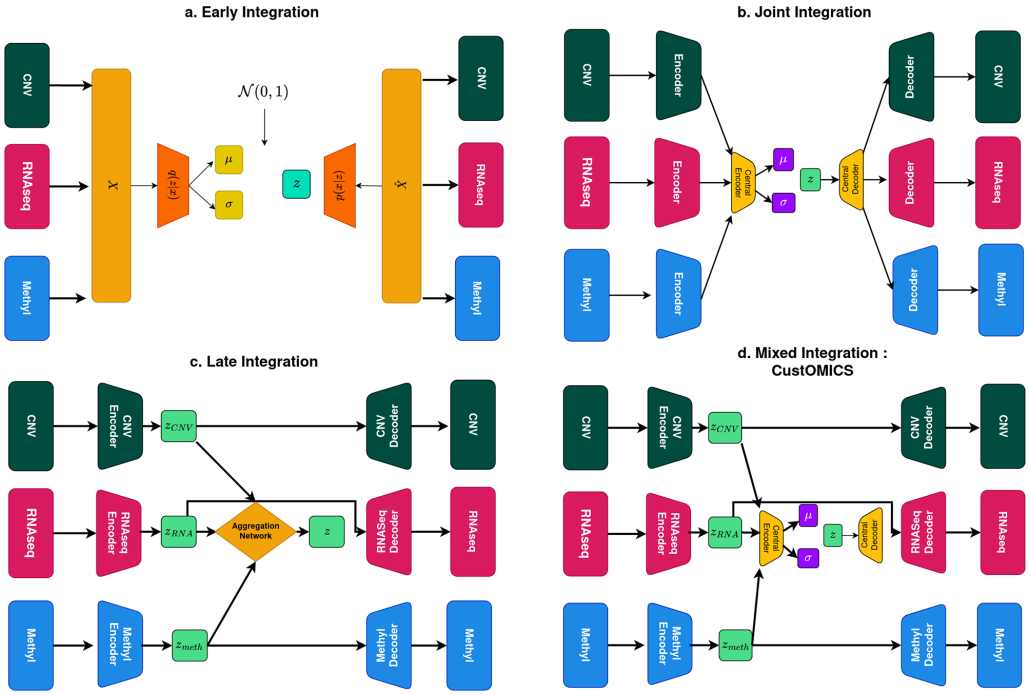


Figure 3.2: *This figure illustrates different integration strategies using Variational Autoencoders (VAEs) for multi-omics data:* ***Early Integration :*** *All omics data (CNV, RNAseq, Methylation) are concatenated into a single input and processed by a shared encoder-decoder network, producing a unified latent representation* z. ***Joint Integration :*** *Each omic modality is encoded separately, and the resulting latent representations are merged in a central encoder-decoder network to learn a shared latent space* z*, capturing both shared and modality-specific features.* ***Late Integration :*** *Modality-specific latent spaces (e.g.,* z_{CNV}*,* z_{RNA}*,* z_{methyl}*) are learned independently and then combined by an aggregation network to form a joint representation* z*, retaining modality-specific information.* ***Mixed Integration (CustOMICS) :*** *Combines phases of independent modality training with frozen central layers (Phase 1) and joint training with unfrozen layers (Phase 2), aiming to balance modality-specific learning with effective multimodal integration.*

We use the deep learning survival framework, DeepSurv, loss function for the survival task. This nonlinear proportional hazard model has been introduced in [133]. The model is built by using the negative partial log-likelihood formula that translates, in our case, into:

$$L(\theta) = -\sum_{i:E_i=1} (\hat{\mu}(x_i;\theta) - log \sum_{j \in \mathcal{R}(T_i)} e^{\hat{\mu}(x_j;\theta)}) \tag{3.1}$$

where E_i is the event for patient I, $\hat{\mu}(x;\theta)$ is the risk function associated with the risk score estimated by the output layer of the network, $\mathcal{R}(t)$ is the risk set, that is the set of patients still at risk of failure after time t.

This task-related loss will also be assigned to each omic-specific network in the first training phase to create adequate sub-representations before the joint integration phase.

Multi-Omics Explainability

To build a more interpretable architecture, we adapt the method introduced in [256] to compute Shapley Additive explanations values, SHAP values [160], for deep variational autoencoders (More details can be found in Appendix A.5). Whereas this method has only been conceived for single-source inputs, we expand it to the multi-source setting of CustOmics by adapting it to any deep autoencoder and applying it to each source autoencoder thanks to the dual-phase approach characterizing the mixed-integration strategy.

After training the CustOmics network, we analyze the importance of features (such as genes) using SHAP (SHapley Additive exPlanations) values. Grounded in cooperative game theory, SHAP values explain the output of machine learning models by attributing the prediction to each input feature.

Each omic modality is trained independently during the first phase of CustOmics training. Specifically, each autoencoder learns a latent representation for its corresponding omic source. Let $X^{(m)} \in \mathbb{R}^{n \times d_m}$ represent the input data matrix for the m-th omic modality, where n is the number of samples and d_m is the number of features (genes) in that modality. The encoder for each modality transforms $X^{(m)}$ into a latent representation $Z^{(m)} \in \mathbb{R}^{n \times k_m}$, where k_m is the dimensionality of the latent space for the m-th modality:

$$Z^{(m)} = f_{\text{enc}}^{(m)}(X^{(m)}; \theta_{\text{enc}}^{(m)})$$

where $\theta_{\text{enc}}^{(m)}$ represents the parameters of the encoder network for modality m.

We then pass the latent representation $Z^{(m)}$ or the prediction output from each autoencoder to a DeepSHAP explainer. This explainer computes the SHAP values $\phi_i^{(m)}$ for each gene i in modality m with respect to a given prediction $\hat{y}$:

$$\phi_i^{(m)} = \text{SHAP}\left(\frac{\partial \hat{y}}{\partial Z^{(m)}}\right)$$

The SHAP values measure the contribution of each gene i in the m-th modality to the prediction. By averaging the SHAP values over a group of samples with similar features, we estimate the overall importance of each gene when considering only that specific omic source.

In the second phase of CustOmics training, the network is jointly trained by unfreezing the central integration layer, allowing the latent representations from all modalities to interact and contribute to the final prediction. The latent vectors from different modalities are combined to form a joint latent space Z, which is used to make the final prediction $\hat{y}$:

$$Z = f_{\text{joint}}\left(Z^{(1)}, Z^{(2)}, \ldots, Z^{(M)}; \theta_{\text{joint}}\right)$$

where M is the total number of omic modalities, and θ_{joint} represents the parameters of the joint network. The combined latent representation Z is used to compute the final prediction.

After this phase, SHAP values $\phi_i^{(m)}$ are computed again, but this time the interactions between different modalities influence them:

$$\phi_i^{(m)} = \text{SHAP}\left(\frac{\partial \hat{y}}{\partial Z}\right)$$

This allows us to observe how the importance of a feature (gene) changes when other modalities are introduced. For example, a gene i in modality m that had a particular importance in Phase 1 may see its importance increase or decrease in Phase 2 due to the additional context provided by other modalities.

By comparing the SHAP values from Phase 1 and Phase 2, we can gain insights into the robustness and interdependence of features across different modalities. The mixed integration approach enables us to assess the standalone importance of genes within a single modality and how these genes' contributions are modulated when other biological layers are considered.

For instance, a gene that was moderately important in the transcriptomic data alone might become highly important when combined with genomic data like CNV, suggesting that the gene's function is closely linked with its number of copys. Alternatively, a highly important gene in isolation may become less important in the joint context, indicating redundancy or interaction effects with other data types.

3.1.2 . Experimental Setup

Test Cases and Datasets

This study uses datasets extracted from the Genomic Data Commons (GDC) pan-cancer multi-omics study [100]. It is one of the most comprehensive datasets for multi-omics analysis, with high-

dimensional omics data and corresponding phenotype data from The Cancer Genome Atlas (TCGA).

Our experiments use three types of omics data: Copy Number Variations (CNV), RNA-Seq gene expressions, and DNA methylation. The CNV dataset comprises Gistic2 measurements on a total of 19,729 genes. The RNA-Seq expression dataset profile comprises around 60,484 identifiers referring to corresponding exons and measuring log2 transformed Fragments Per Kilobase of transcript per Million mapped reads, FPKM. Finally, the DNA methylation dataset was produced using the Infinium HumanMethylation450 BeadChip (450K) arrays with 485,578 probes in which beta values of probes indicate the methylation ratio of corresponding CpG sites.

Moreover, we also evaluate our method on five smaller cohorts from TCGA: Bladder Urothelial Carcinoma (BLCA, n=437), Breast Invasive Carcinoma (BRCA, n=1022), Lung Adenocarcinoma (LUAD, n=498), Glioblastoma & Lower Grade Glioma (GBMLGG, n=515) and Uterine Corpus Endometrial Carcinoma (UCEC, n=538). Appendix B.1 shows more details on the different datasets used.

We will perform in this study 4 different evaluations on 4 test cases for classification and survival. The first task in our set of experiments is classifying the different tumor types in the pan-cancer study. The classification performance was measured using five metrics: Accuracy, Macro-averaged F1 score, precision, recall, and ROC-AUC [259]. We also perform a second classification task to validate our findings on a smaller dataset and test the robustness of our method. This task aims to perform a tumor subtype classification based on the PAM50 classification (LuminalA, LuminalB, Basal, and HER2). We use the same setup as the pan-cancer case. The third test case will be a survival study of the Pancancer dataset. Finally, the fourth test case will evaluate the survival performances of the five datasets presented earlier. For all those test cases, We compare the CustOmics model to several reference methods for multi-omics integration: first with a combination of dimensionality reduction methods, Multiple Factor Analysis, MFA [215], Uniform Manifold Approximation and Projection, UMAP [170], and non-negative matrix factorization, NMF [272], and also with different deep learning methods corresponding to the various strategies described in the first chapter.

Data Preprocessing

The data required some preprocessing before analysis.

- For RNA gene expression profiles, 594 exons located on the Y chromosome were removed to avoid sex-specific expression, along with 1,904 ones with zero expression and 248 with missing values.

- For DNA methylation data, the same strategy as with gene expression profiles was used, in addition to removing probes that cannot be mapped to the human reference genome. It leaves us with 438,831 CpG sites.

We then intersected the samples across all combinations of omics data to maximize the number of samples available for each test case. After this, we identified and removed any features that had missing values, consistently zero values, or NA entries across the other omic datasets. Finally, to ensure that each omic source was equally weighted during integration, we applied min-max normalization to the non-normalized datasets, such as CNVs and RNA-Seq data, so that all omic data sources were scaled consistently.

Implementation Details

The CustOmics framework is based on the Pytorch deep-learning library [192]. It can be applied to any combination of high-dimensional datasets and incorporates different integration strategies depending on the type of data and task to perform. As done in *Zhang et al.* [273], DNA methylation data can be divided into 23 separate blocks, each feeding a hidden layer corresponding to a chromosome to avoid overfitting and save GPU memory.

The whole architecture is built using fully connected blocks with weights initialized following a uniform distribution $\mathcal{U}(-\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{k}})$ where k is the number of weight parameters. We use a batch normalization technique in each layer composing the neural network to address the internal covariate shift problem[121]. Also, to avoid overfitting problems, we use dropout [228]; its rate is considered a hyperparameter.

The input dataset was randomly split into training, validation, and testing sets (60-20-20%) using stratified 5-fold cross-validation so that the proportion of samples in each tumor type between the different sets is preserved in all the folds. We perform Bayesian optimization [227] using the validation set to find our model's best possible combination of hyperparameters.

All models were trained using an Nvidia Tesla V100S with 32 GB memory.

The CustOmics framework is open-source and available in Github.

3.1.3 . Results

Classification Results

We first perform the classification task on the pan-cancer dataset. Each architecture is coupled with an artificial neural network classifier composed of two hidden layers with 256 and 128 neurons. This network is trained using a categorical cross-entropy loss with ReLU activation function on the hidden layer and a Softmax activation function on the output layer.

Table 3.1: *The classification performance for the pan-cancer dataset is evaluated with 5 standard metrics for UMAP, NMF, MFA, Unsupervised Customics with SVM, and supervised deep-learning methods. We evaluate the performances on the final predicted output of the downstream classifier. Best results are in bold.*

Model	Accuracy	F1-score	Precision	Recall	ROC-AUC
UMAP	0.7598 ± 0.0036	0.7149 ± 0.0029	0.7200 ± 0.0031	0.7598 ± 0.0032	0.8740 ± 0.0012
NMF	0.8599 ± 0.0017	0.8406 ± 0.0013	0.8460 ± 0.0018	0.8599 ± 0.0021	0.9266 ± 0.0019
MFA	0.9167 ± 0.0012	0.9025 ± 0.0014	0.8945 ± 0.0008	0.9167 ± 0.0013	0.9565 ± 0.0003
Unsup. Cust.	0.9335 ± 0.0038	0.9323 ± 0.0039	0.9342 ± 0.0043	0.9335 ± 0.0038	0.9689 ± 0.0019
Early Int. VAE	0.9337 ± 0.0079	0.9314 ± 0.0086	0.9367 ± 0.0067	0.9337 ± 0.0079	0.9655 ± 0.0041
Joint Int. VAE	0.9610 ± 0.0032	0.9600 ± 0.0032	0.9631 ± 0.0043	0.9610 ± 0.0032	0.9898 ± 0.0005
Late Int. VAE	0.9492 ± 0.0115	0.9464 ± 0.0111	0.9498 ± 0.0079	0.9492 ± 0.0115	0.9737 ± 0.0060
Mix Int. AE	0.9453 ± 0.0056	0.9423 ± 0.0063	0.9452 ± 0.0050	0.9453 ± 0.0056	0.9717 ± 0.0029
CustOmics	$\mathbf{0.9788 \pm 0.0025}$	$\mathbf{0.9705 \pm 0.0033}$	$\mathbf{0.9728 \pm 0.0041}$	$\mathbf{0.9685 \pm 0.0034}$	$\mathbf{0.9918 \pm 0.0001}$

Table 3.2: *Classification performances for multiple combinations of omics data using joint integration on the pan-cancer dataset.*

Omics	Accuracy	F1-score	Precision	Recall	ROC-AUC
CNV	0.47 ± 0.03	0.47 ± 0.03	0.47 ± 0.03	0.48 ± 0.01	0.75 ± 0.02
RNAseq	0.92 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	0.92 ± 0.01	0.96 ± 0.00
methyl	0.68 ± 0.02	0.68 ± 0.02	0.68 ± 0.02	0.68 ± 0.02	0.82 ± 0.01
CNV + RNAseq	0.90 ± 0.02	0.89 ± 0.02	0.90 ± 0.02	0.88 ± 0.03	0.93 ± 0.00
CNV + methyl	0.70 ± 0.02	0.69 ± 0.02	0.70 ± 0.02	0.70 ± 0.02	0.85 ± 0.01
RNAseq + methyl	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.96 \pm 0.01}$	0.96 ± 0.0132	$\mathbf{0.99 \pm 0.00}$
CNV + RNAseq + methyl	0.94 ± 0.01	0.94 ± 0.01	0.95 ± 0.01	0.94 ± 0.01	0.97 ± 0.00

Table 3.3: *Classification performances for multiple combinations of omics data using CustOMICS on the pan-cancer dataset. We can see that RNAseq data bring the best performances, but adding other omics data increases the performances, suggesting that the integration is relevant.*

Omics	Accuracy	F1-score	Precision	Recall	ROC-AUC
CNV	0.47 ± 0.03	0.47 ± 0.03	0.47 ± 0.03	0.48 ± 0.01	0.75 ± 0.02
RNAseq	0.92 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	0.92 ± 0.01	0.96 ± 0.00
methyl	0.68 ± 0.02	0.68 ± 0.02	0.68 ± 0.02	0.68 ± 0.02	0.82 ± 0.01
CNV + RNAseq	0.93 ± 0.01	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01	0.96 ± 0.00
CNV + methyl	0.71 ± 0.02	0.69 ± 0.02	0.70 ± 0.02	0.69 ± 0.02	0.85 ± 0.01
RNAseq + methyl	0.94 ± 0.01	0.94 ± 0.01	0.94 ± 0.01	0.94 ± 0.0132	0.97 ± 0.00
CNV + RNAseq + methyl	$\mathbf{0.98 \pm 0.01}$	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.99 \pm 0.00}$

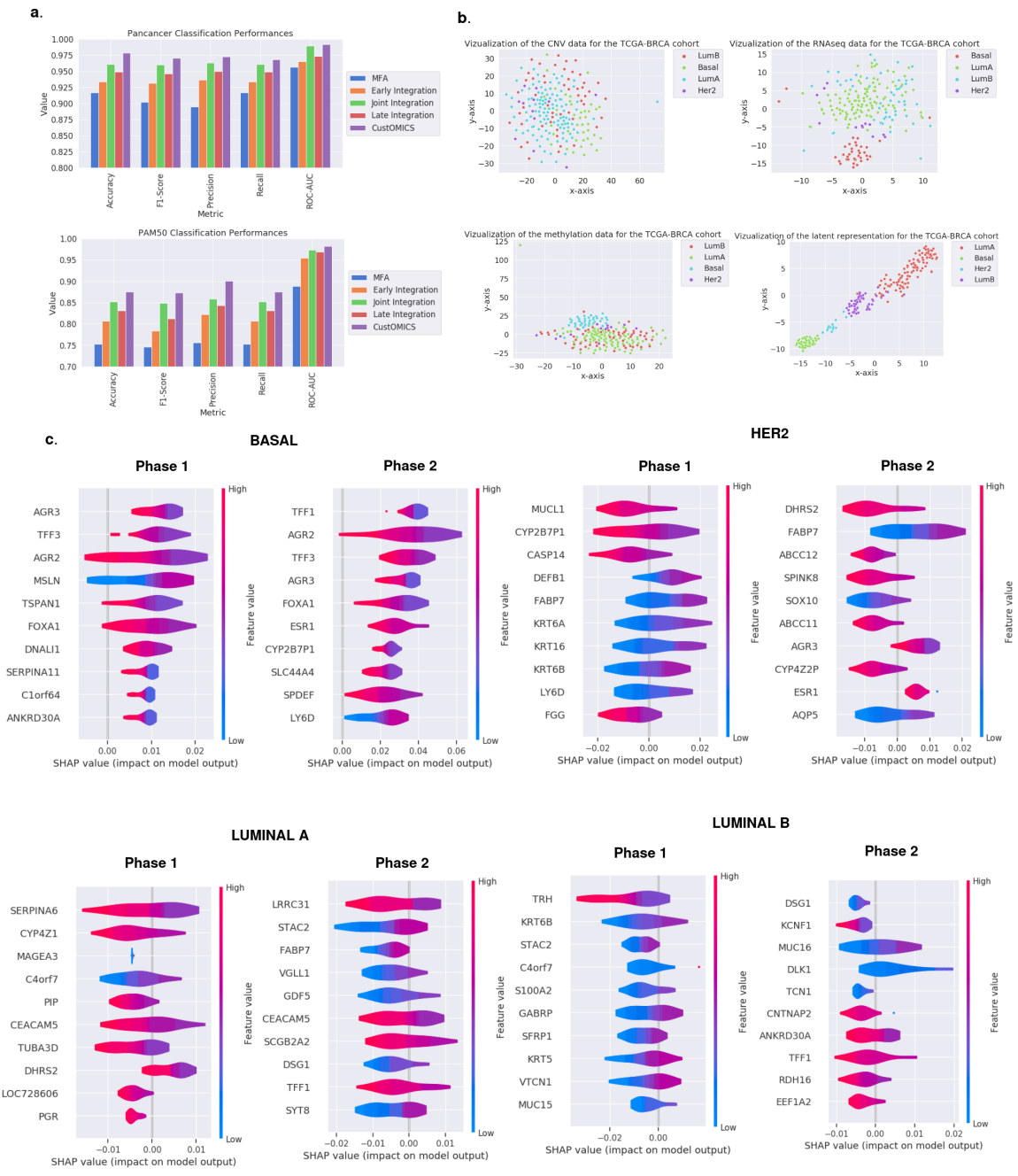


Figure 3.3: ***a. pan-cancer and PAM50 classification results:*** *Overall classification results for the pan-cancer tumor classification test case and the PAM50 subtype classification for breast cancer.* ***b. T-SNE vizualization*** *for each omic source separately, along with the latent representation constructed by CustOmics. We see that the constructed layer representation succeeds at separating the data into four distinct clusters that we couldn't distinguish with each omic source alone.* ***c. PAM50 gene importance:*** *Computed SHAP values on the RNA-Seq data of the most relevant genes responsible for discriminating between subtypes against the others using CustOmics for both integration phases.*

Fig 3.3, Table 3.1 and Table 3.4 show the overall classification results.

Among the factorial methods, the MFA achieved the best results, so we coupled this method with the same ANN classifier used with the deep-learning representation methods as a basis for comparison. However, it does not perform as well as most deep-learning methods. This is because MFA cannot uncover nonlinear relationships between different sources, unlike deep-learning architectures. Moreover, as the MFA is an early integration, it suffers from problems of signal imbalance related to early integration.

We also assessed the performance of an unsupervised representation given by our CustOmics network by plugging it into the same ANN classifier as used for the factorial methods. This unsupervised setting performs quite well compared to similar unsupervised methods like the factorial ones, which show the robustness of the representation learned by CustOmics, even without adding the task loss. Moreover, we can see in Table 3.1 that, generally, Variational Autoencoders perform better than standard autoencoders, which comforts us in choosing a variational setup for CustOmics.

As hinted earlier, we can see that for the deep learning strategies, early integration is behind the others in terms of performance. It can be explained by the fact that RNA-Seq data hold more signals when determining tumor types or subtypes. Thus, concatenating the sources before feeding them to the VAE overshadows the other sources, and the learned representation depends mostly on RNA-Seq data without leveraging the other modalities. It is illustrated in Table 3.3, showing that the classification results using RNA-Seq data only are very close to those obtained with early integration, indicating that the model may overlook the interactions between sources. Late integration is not optimal since interactions between sources are not adequately learned as the predictors are trained separately. Joint integration performs well in most cases, but we see that the best results displayed in 3.2 are achieved by the combination of only two sources, RNA-Seq and Methylation data, as it seems that CNV data only adds noise to the latent representation, meaning that its information is not handled well with this strategy.

These results confirm the interest in the CustOmics architecture, as it gives the best performances for all the test cases while being able to converge without apparent overfitting, as suggested by Figure C.2. Moreover, it also takes advantage of the complementarity and interactions between sources: As shown in Appendix B3, all sources bring additional information. We can witness that even though transcriptomics data gives the best performances, other omics sources succeed at bringing additional

Table 3.4: *Classification performance for PAM50 breast cancer subtype classification on the TCGA-BRCA dataset, with 5 standard metrics. We compare machine-learning methods like UMAP, NMF, and MFA with deep-learning methods. We evaluate the performances on the final predicted output of the downstream classifier. Best results are in bold.*

Model	Accuracy	F1-score	Precision	Recall	ROC-AUC
UMAP	0.6815 ± 0.0152	0.6612 ± 0.0140	0.6637 ± 0.0157	0.6815 ± 0.0151	0.8482 ± 0.0055
NMF	0.7025 ± 0.0132	0.6955 ± 0.0135	0.7032 ± 0.0110	0.7025 ± 0.0131	0.8576 ± 0.0119
MFA	0.7532 ± 0.0164	0.7460 ± 0.0160	0.7562 ± 0.0162	0.7531 ± 0.0165	0.8884 ± 0.0037
Unsup. Cust.	0.8133 ± 0.0172	0.8044 ± 0.0168	0.8322 ± 0.0144	0.8112 ± 0.0150	0.9674 ± 0.0031
Early Int. VAE	0.8063 ± 0.0152	0.7840 ± 0.0167	0.8228 ± 0.0167	0.8063 ± 0.0150	0.9552 ± 0.0077
Joint Int. VAE	0.8518 ± 0.0184	0.8488 ± 0.0189	0.8589 ± 0.0161	0.8518 ± 0.0151	0.9734 ± 0.0035
Late Int. VAE	0.8312 ± 0.0174	0.8124 ± 0.0201	0.8429 ± 0.0215	0.8312 ± 0.0176	0.9689 ± 0.0066
Mix Int. AE	0.8452 ± 0.0168	0.8322 ± 0.0214	0.8477 ± 0.0234	0.8417 ± 0.0181	0.9709 ± 0.0051
CustOmics	$\mathbf{0.8758 \pm 0.0162}$	$\mathbf{0.8728 \pm 0.0141}$	$\mathbf{0.9012 \pm 0.0137}$	$\mathbf{0.8758 \pm 0.0130}$	$\mathbf{0.9828 \pm 0.0022}$

information. The high performance of transcriptomics data is predictable as most information about tumor types and molecular subtypes is expressed in RNA data. It is also interesting to see that in Appendix B1, transcriptomics data need not only a few layers to converge to their best results, but other data types do not.

Regarding the computational cost of all those methods, we see in Table C.1 that they all have around the same number of trainable parameters. The slight increase in the number of parameters in CustOmics is due to the intermediate networks necessary for phase 1, similar to the late integration setup. Figure 3.3 gives a visualization of the different sources: Even though the initial sources are quite entangled, the CustOmics latent representation separates the clusters using the mutual information between modalities.

We also use the interpretability property of CustOmics introduced in the Methods section to highlight the most relevant features for discriminating between PAM50 subtypes by computing their respective SHAP values for each source. We do it for both phases: in phase 1, we retrieve the relevant genes considered when using a single omic source, whereas, in phase 2, we investigate how adding other sources' signals changes the genes' importance. Fig 3.3 references the results of such explanations on RNA-Seq data, Figure C.3 and Fig C.4 show the results for CNV and methylation data. We observe some well-referenced biomarkers for breast cancer like TFF1 [264], suggesting that our method can retrieve relevant biological information.

Survival Analysis

The second task in this study is survival analysis, where the objective is to predict the risk score associated with each patient based on the corresponding high-dimensional omics data. To evaluate the performance of this downstream task, two standard metrics are used: the Concordance Index (C-index), which generalizes the AUC metric for censored survival data [110], and the Integrated Brier Score (IBS) [95].

For the Pancancer analysis, where different cancer types have varying weights and mortality rates, we implemented a weighted C-index to ensure a fair evaluation across all tumor types. In this approach, the trained model is first evaluated on each tumor type individually to compute the C-index for that specific type. We then perform a weighted average of these C-indices, with the weights reflecting each cancer type's relative importance or prevalence. This method ensures that the overall performance metric fairly represents the model's effectiveness across diverse tumor types.

Fig 3.4, Table 3.5, and Table 3.7 show the results for the different methods for the survival task. The same observations regarding the differences between integration strategies can be made regarding the classification task. Here again, we also evaluated the performance of the CustOmics method for each combination of omics sources as shown in Table 3.6.

Table 3.5: *The survival analysis performance for the pan-cancer dataset is evaluated with two standard metrics, C-index and IBS. We compare classical methods like UMAP, NMF, and MFA with deep-learning methods and evaluate the performances on the final predicted output of the downstream survival network. Best results are in bold.*

Model	C-index	IBS
UMAP	0.5948 ± 0.0231	0.2486 ± 0.0327
NMF	0.6012 ± 0.0204	0.2207 ± 0.0264
MFA	0.6127 ± 0.0164	0.2192 ± 0.0203
Unsup. Cust.	0.6329 ± 0.0144	0.2087 ± 0.0207
Early Int. VAE	0.6578 ± 0.0103	0.2106 ± 0.0117
Joint Int. VAE	0.6709 ± 0.0041	0.1802 ± 0.0072
Late Int. VAE	0.6629 ± 0.0086	0.2112 ± 0.0088
Mix Int. AE	0.6618 ± 0.0051	0.1815 ± 0.0074
CustOmics	$\mathbf{0.6841 \pm 0.0033}$	$\mathbf{0.1745 \pm 0.0052}$

The last task consists in evaluating the model performances for survival analysis for several specific cancer types of the TCGA datasets described in the dataset section. The objective is to evaluate the robustness of the models when dealing with smaller datasets.

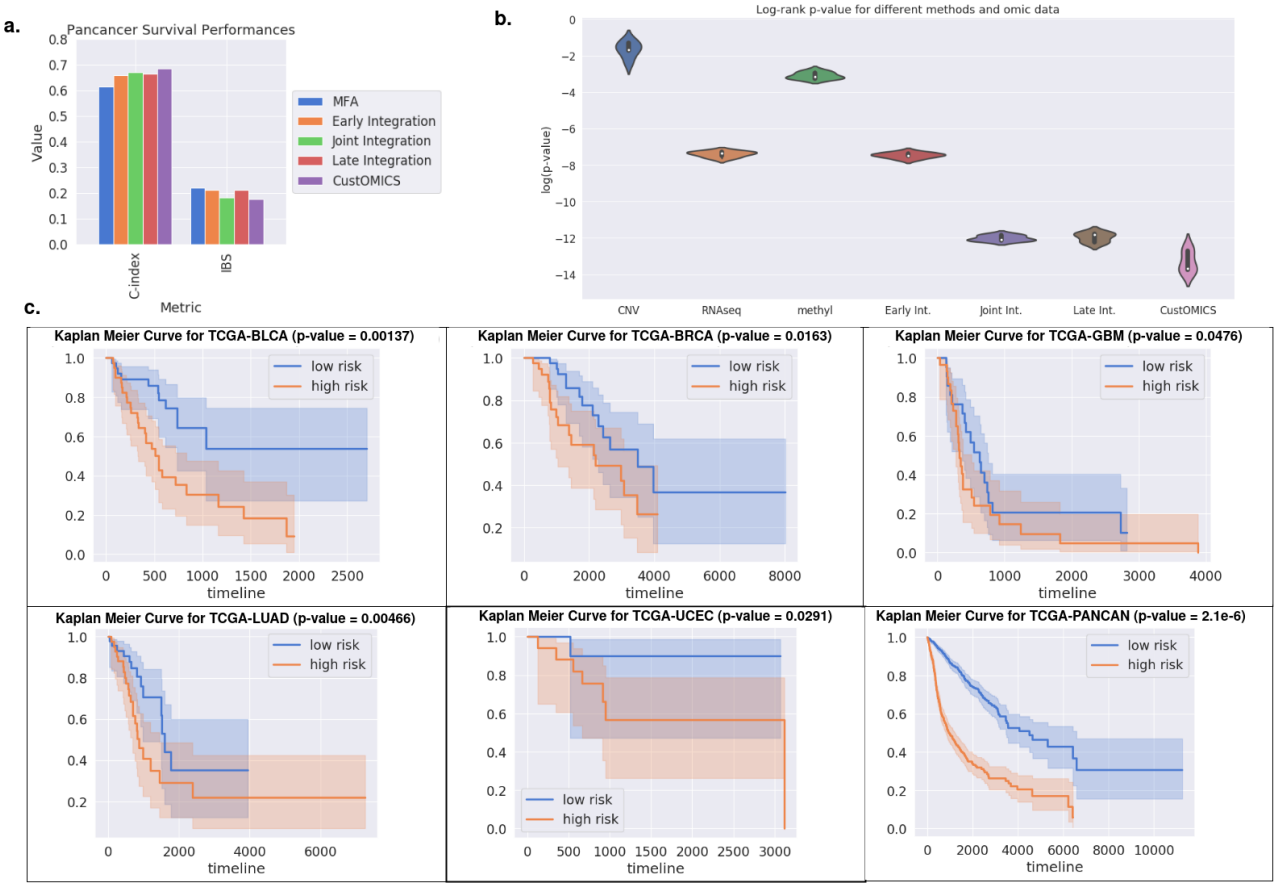


Figure 3.4: ***a. Survival Analysis Performances:*** *We evaluate the performances of the survival model for the pan-cancer dataset using both the C-index and the Integrated Brier Score (IBS). Here again, our model outperforms the other integration strategies on both metrics.* ***b. Log-rank test:*** *We compute the p-value associated with the log-rank test between high and low-risk groups for every integration strategy on a validation set for the pan-cancer survival test case and compare it to mono-omic survival predictions.* ***c. Kaplan Meier Curves :*** *We draw the Kaplan Meier curves and display the p-value associated with the log-rank test as computed previously for each dataset using the predicted hazard from the CustOmics model and stratify the population into high and low risk on the test set for the predicted hazard ratio. This figure shows that our method successfully stratifies the patients into risk subgroups.*

Finally, we perform a more thorough analysis of the survival results we display in Fig 3.4. We leave out 20% of our datasets for validation purposes, and we perform 5-fold cross-validation on the remaining 80% to compute the p-values associated with the log-rank test for different combinations of the Pan-cancer test case. We show the ability of CustOmics to stratify the patients into distinct risk groups using the predicted hazard ratio. This stratification ability was also measured quantitatively using the p-value associated with the log-rank test between the different categories. Even though the

Table 3.6: *Survival performances for multiple combinations of omics data using CustOMICS on the pan-cancer dataset. We can see that the best performances are obtained with RNAseq data, but the addition of other omics data increases the performances, suggesting that the integration is relevant.*

Omics	Accuracy	F1-score
CNV	0.54 ± 0.05	0.25 ± 0.06
RNASeq	0.63 ± 0.02	0.20 ± 0.02
methyl	0.59 ± 0.02	0.23 ± 0.03
CNV + RNAseq	0.64 ± 0.02	0.19 ± 0.02
CNV + methyl	0.61 ± 0.03	0.21 ± 0.03
RNAseq + methyl	0.64 ± 0.02	0.19 ± 0.02
CNV + RNAseq + methyl	$\mathbf{0.68 \pm 0.01}$	$\mathbf{0.17 \pm 0.01}$

Table 3.7: *Survival performances of state-of-the-art integration methods for survival analysis, using concordance index on 5 TCGA cohorts: Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Glioblastoma & Lower Grade Glioma (GBMLGG), Lung Adenocarcinoma (LUAD) and Ulterine Corpus Endometrial Carcinoma (UCEC).*

Model	BLCA	BRCA	GBMLGG	LUAD	UCEC	Overall
UMAP	0.527 ± 0.048	0.524 ± 0.039	0.557 ± 0.028	0.530 ± 0.017	0.543 ± 0.025	0.536
NMF	0.553 ± 0.060	0.560 ± 0.036	0.584 ± 0.067	0.589 ± 0.052	0.570 ± 0.068	0.571
MFA	0.591 ± 0.052	0.599 ± 0.043	0.605 ± 0.036	0.597 ± 0.058	0.586 ± 0.020	0.596
Unsup. Cust.	0.599 ± 0.062	0.624 ± 0.048	0.633 ± 0.032	0.606 ± 0.049	0.622 ± 0.023	0.617
Early Int. VAE	0.603 ± 0.054	0.618 ± 0.039	0.628 ± 0.056	0.612 ± 0.041	0.609 ± 0.032	0.614
Joint Int. VAE	0.616 ± 0.072	0.627 ± 0.030	0.635 ± 0.020	0.608 ± 0.038	0.630 ± 0.021	0.624
Late Int. VAE	0.610 ± 0.055	0.620 ± 0.057	0.621 ± 0.067	0.595 ± 0.012	0.627 ± 0.022	0.615
Mix Int. AE	0.611 ± 0.049	0.625 ± 0.061	0.609 ± 0.047	0.594 ± 0.031	0.615 ± 0.020	0.611
CustOmics	$\mathbf{0.637 \pm 0.050}$	$\mathbf{0.633 \pm 0.018}$	$\mathbf{0.642 \pm 0.028}$	$\mathbf{0.625 \pm 0.037}$	$\mathbf{0.667 \pm 0.022}$	$\mathbf{0.640}$

comparison between joint, late, and mixed integration is not evident in this case, it is interesting to note that the addition of multiple omics sources has dramatically affected the p-value as it was nearly multiplied by a factor 10^{-5} for CNV data and 10^{-2} for RNA-Seq and methylation data. Those results also show that early integration is the strategy with minimal enrichment from other sources. This corroborates our previous intuition and is coherent with the results found in the different experiments.

3.2 . An Application of Multi-Omics Integration to Myelodysplastic Syndromes

Myelodysplastic syndromes (MDS) are diverse hematopoietic disorders characterized by ineffective blood cell production, resulting in low blood cell counts and an increased chance of developing acute myeloid leukemia (AML). The development and function of bone marrow stem cells in MDS are influenced by various genetic, epigenetic, and environmental factors, making the pathogenesis com-

plex. In a clinical setting, diagnosing and treating MDS is difficult because of its various symptoms and results.

The primary basis for diagnosing MDS is examining the bone marrow and peripheral blood morphology, along with cytogenetic studies. However, these customary approaches frequently need to capture the complex molecular makeup of MDS. The diversity of genetic changes in MDS patients is highlighted by gene mutations involved in RNA splicing, DNA methylation, and histone modification. The diverse characteristics of these changes make the prediction and treatment planning more difficult, underscoring the urgent need for enhanced diagnostic instruments.

3.2.1 . Context

Myelodysplastic Syndromes (MDS) represent a heterogeneous group of hematopoietic stem cell disorders characterized by ineffective hematopoiesis, dysplasia in one or more of the significant myeloid cell lines, and an increased risk of progression to acute myeloid leukemia (AML). Over the past few decades, substantial research has been conducted to understand the pathophysiology, classification, diagnosis, and treatment of MDS.

The pathophysiology of MDS involves a complex interplay of genetic and epigenetic alterations. Mutations in genes such as *TP53*, *RUNX1*, *ASXL1*, and *SF3B1* have been identified as key drivers in MDS development and progression. These mutations often affect crucial cellular processes, including DNA methylation, histone modification, and RNA splicing, leading to aberrant hematopoiesis and clonal evolution [190]. Recent studies using next-generation sequencing (NGS) have provided more profound insights into the mutational landscape of MDS, revealing that multiple mutations are often present in a hierarchical order, contributing to the disease's heterogeneity [104].

The World Health Organization (WHO) classification system for MDS, updated in 2016, emphasizes the importance of cytogenetic and molecular abnormalities in addition to morphological criteria. This classification distinguishes MDS from related myeloid neoplasms such as chronic myelomonocytic leukemia (CMML), which has myelodysplastic and myeloproliferative features [10]. The revised International Prognostic Scoring System (IPSS-R) incorporates cytogenetic findings and clinical parameters to stratify patients based on risk, aiding prognosis and treatment decisions [97].

Integrating multi-omics approaches, including genomics, transcriptomics, and epigenomics, has revolutionized the diagnosis and understanding of MDS. Copy number variations (CNVs), variant allele frequencies (VAFs), and RNA sequencing (RNA-seq) provide comprehensive molecular profiles that

enhance disease classification and reveal potential therapeutic targets. For instance, the application of NGS and single-cell RNA-seq has uncovered clonal architecture and evolutionary dynamics, offering insights into disease progression and resistance mechanisms [70].

Therapeutic strategies for MDS have evolved significantly, with options ranging from supportive care to disease-modifying treatments. Hypomethylating agents (HMAs) such as azacitidine and decitabine have improved survival and delayed progression to AML [79]. Additionally, lenalidomide has been effective in MDS with del(5q) cytogenetic abnormalities [152]. Allogeneic hematopoietic stem cell transplantation (HSCT) remains the only curative option, though it is limited by patient age and comorbidities [63]. The development of targeted therapies, including inhibitors of mutant proteins and immune checkpoint inhibitors, holds promise for more personalized treatment approaches [270].

Identifying reliable biomarkers for prognosis and treatment response is a critical area of research in MDS. Studies have shown that specific gene mutations and cytogenetic abnormalities are associated with distinct clinical outcomes. For example, mutations in *TP53* are linked to poor prognosis, while mutations in *SF3B1* often indicate a more favorable outcome [22]. Comprehensive mutational profiling in clinical practice can guide therapeutic decisions and risk stratification.

The Molecular International Prognostic Scoring System (IPSS-M), developed by Elsa Bernard et al. [26], enhances the risk stratification of Myelodysplastic Syndromes (MDS) by integrating molecular genetic data with traditional clinical and cytogenetic parameters. This system incorporates mutations from 31 essential genes, providing a more accurate and personalized risk assessment. Compared to the previous IPSS-R, the IPSS-M reclassified 46% of patients into different risk categories, improving prognostic precision and treatment decision-making. The IPSS-M also addresses therapy-related MDS, offering a versatile and comprehensive tool for guiding patient care and clinical trial design.

Despite these advances, challenges still need to be addressed in managing MDS. The heterogeneous nature of the disease, variability in patient responses, and the development of resistance to existing therapies highlight the need for ongoing research. Future directions include integrating advanced multi-omics techniques to uncover novel therapeutic targets, developing more effective combination therapies, and exploring the tumor microenvironment's role in disease progression [171].

This study will focus on assessing the use and performance of a multi-omics analysis for Myelodysplastic Syndromes.

3.2.2 . Data Description & Preprocessing

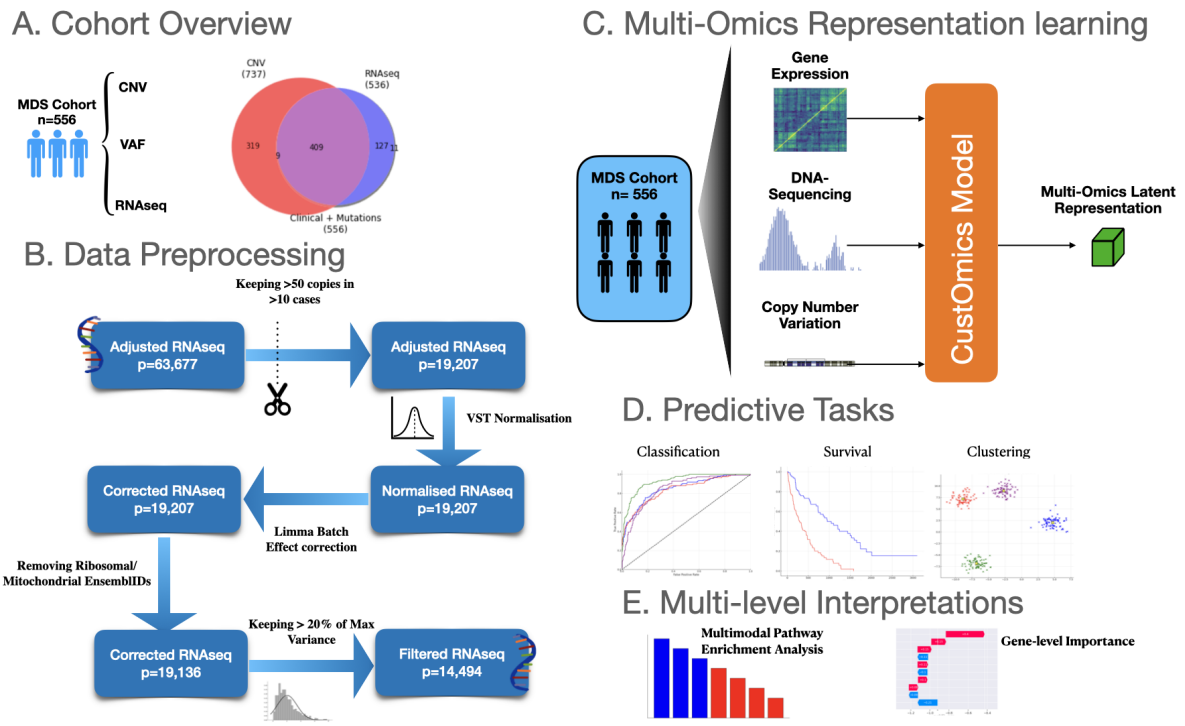


Figure 3.5: ***A. Cohort Overview:*** *Overview of the MDS cohort with the available molecular sources (VAF, CNV, RNAseq).* ***B. Data Processing:*** *Processing Pipeline for the RNAseq Data.* ***C. Mulit-Omics Representation learning:*** *Use of the CustOmics Model to create a latent representation of multi-omics data.* ***D. Predictive Tasks:*** *For this study, we will perform classification, survival and unsupervised clustering tasks.* ***E. Multi-Level Interpretations:*** *CustOmics offers explainability at different levels of the biological system.*

The dataset includes information on clinical and molecular aspects of 556 MDS patients with Copy Number Variations (CNV), Gene expression & VAFs (More details can be found in Appendix B.2). Patients were obtained from three distinct centers, providing a wide range of the disease. The group consists of male and female patients, with slightly more males.

The data contains various hematologic measurements, such as the Revised International Prognostic Scoring System (IPSS-R) factors like blast percentages and hemoglobin levels. These factors are essential for evaluating the seriousness and advancement of the illness.

The WHO 2016 criteria categorize patients into different subgroups, including various MDS subtypes. MDS with multilineage dysplasia (MDS_MLD) and MDS with ring sideroblasts and multilineage dysplasia (MDS_RS_MLD) are prevalent subtypes that demonstrate the various and intricate nature of the disease in this group.

Data on survival, such as overall survival (OS) and event-free survival (EFS) times, offer information

on patient outcomes and treatment effectiveness. In addition, information on using erythropoiesis-stimulating agents (EPO) and azacitidine (AZA) is provided, focusing on treatment methods and how long they were used in this group.

Initially, the dataset of 63,677 adjusted RNAseq reads undergoes a filtering process to retain Ensembl IDs with more than 50 copies in at least 10 cases, resulting in 19,207 genes. This step is crucial for excluding low-abundance transcripts and reducing noise. Following this, the data is normalized using Variance Stabilizing Transformation (VST), stabilizing the variance across different expression levels, making the dataset more comparable across samples while maintaining the same number of genes.

To address potential technical variations arising from different experimental conditions, batch effect correction on the center variable is performed using the Limma method [202]. This correction ensures that the observed gene expression differences are biological rather than technical artifacts, refining the dataset to 19,136 genes. Additionally, ribosomal and mitochondrial genes, which can dominate the RNAseq data and obscure relevant biological signals, are removed, maintaining focus on nuclear gene expression and keeping the gene count consistent.

The final preprocessing step involves filtering the remaining genes based on their variance, retaining only those with a variance greater than 0.5. This step further refines the dataset to 14,494 genes, ensuring that only genes with significant variability and potential biological relevance are included in the final processed dataset.

3.2.3 . Methods

This study employed the CustOmics framework for supervised and unsupervised tasks, as outlined previously.

Unsupervised Clustering

For unsupervised clustering, we adopted a methodology inspired by DeepCluster [40], utilizing self-supervised learning to improve the clustering of multi-omics data. The process begins with the initial clustering of integrated omics data using k-means, a widely used algorithm for partitioning data into k clusters. Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$ where each x_i represents a multi-omics data point, k-means aims to minimize the within-cluster sum of squares (WCSS) by solving:

$$\min_{C_1,\ldots,C_k} \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

where C_j denotes the j-th cluster and μ_j is the centroid of C_j.

These initial clusters serve as the foundation for training the CustOmics framework, which is designed to learn a latent space representation Z of the data. The latent space representation Z is parameterized by a neural network $f_\theta : X \to Z$, where θ represents the learnable parameters of the network. The goal is to find a representation Z such that when the data is re-clustered in this space, the clusters are more compact and distinct.

The training process is iterative, involving the following steps:

1. **Clustering Step:** Given the current latent space representation $Z = f_\theta(X)$, apply k-means clustering to Z, resulting in new cluster assignments $\{C_1, C_2, \ldots, C_k\}$.

2. **Feature Learning Step:** Using the cluster assignments as pseudo-labels, update the parameters θ of the neural network by minimizing a self-supervised loss function $\mathcal{L}$. A common choice for $\mathcal{L}$ is a contrastive loss or a cross-entropy loss defined as:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} \log P(c_i|x_i;\theta)$$

where c_i is the pseudo-label (cluster assignment) for data point x_i, and $P(c_i|x_i;\theta)$ is the probability of assigning x_i to cluster c_i given by the neural network.

3. **Latent Space Update:** After updating θ, recompute the latent space representation $Z = f_\theta(X)$.

4. **Iteration:** Repeat the clustering step with the updated Z and iterate the process until convergence, typically when the change in cluster assignments or the loss $\mathcal{L}$ becomes negligible.

Through this iterative refinement, the CustOmics framework enhances the clustering by ensuring that the latent space representation Z evolves to better reflect the underlying structure of the multi-omics data. The objective is to reduce the intra-cluster variance and increase inter-cluster separation, which is formally expressed as:

$$\min_{\theta} \sum_{j=1}^{k} \sum_{z_i \in C_j} \|z_i - \mu_j^Z\|^2 + \lambda \mathcal{L}(\theta)$$

where μ_j^Z is the centroid of cluster C_j in the latent space, and λ is a regularization parameter.

Ultimately, these refined latent space representations are used to classify data points into distinct clusters, revealing underlying biological patterns that might be missed by supervised methods alone. This approach provides a robust framework for interpreting complex multi-omics datasets and uncovering novel biological insights, consistent with other applications of self-supervised learning in omics data analysis.

Pathway Analysis

Additionally, to further enhance the interpretability of our results, we utilized SHAP values derived from the CustOmics framework as the basis for Gene-Set Variation Analysis (GSVA). SHAP (SHapley Additive exPlanations) values, denoted as $\phi_{i,j}$, quantify the contribution of each feature j to the prediction for a particular sample i. These values are computed based on Shapley values from cooperative game theory, ensuring a fair allocation of importance among the features.

Inspired by the work of Lundberg et al. [160], we replaced the traditional GSVA scores with individual SHAP values. Given a gene set S consisting of genes $g_1, g_2, \ldots, g_m$, the traditional GSVA method calculates an enrichment score ES_S for each gene set S and each sample i, which reflects the extent to which the genes in S are coordinately up- or down-regulated.

In our modified approach, we calculate a SHAP-based enrichment score ES_S^{SHAP} for each gene set S and each sample i as follows:

$$ES_S^{\text{SHAP}}(i) = \sum_{j \in S} \phi_{i,j}$$

where $\phi_{i,j}$ is the SHAP value for gene j in sample i. This score represents the cumulative contribution of the genes in set S to the model's prediction for sample i.

This integration of SHAP values into GSVA provides a nuanced understanding of gene-set variation, enhancing the explainability of our findings. By utilizing SHAP values instead of traditional expression values, we can directly interpret the impact of gene sets on model predictions, thereby offering a more transparent and interpretable analysis of gene-set activity in the context of the multi-omics data.

3.2.4 . Experimental Setup

To evaluate the effectiveness of our proposed method, we conducted three distinct experiments: subtype classification, survival analysis, and unsupervised clustering. Each experiment is designed to assess different aspects of the model's performance.

The first experiment involves a binary classification task aimed at distinguishing between two specific subtypes within the dataset (Details on the dataset are available in Appendix B.2). This task is performed using a 5-fold cross-validation approach to ensure the robustness and generalizability of the results. The dataset is split into five folds, where in each iteration, four folds are used for training, and the remaining fold is used for testing. The process is repeated five times, and the final performance metrics are averaged across all folds.

The evaluation metrics for this classification task include balanced accuracy, F1-score, Precision-Recall, and ROC-AUC. These metrics provide a comprehensive assessment of the model's ability to accurately classify the subtypes, particularly in scenarios where class imbalance may be present. Our method is primarily compared to a penalized logistic regression model, which serves as a baseline. Additionally, we benchmark our approach against several state-of-the-art methods. Detailed results and further analysis of this experiment can be found in the Appendix C.2.2.

The second experiment focuses on evaluating the Leukemia Free Survival (LFS) prediction. For this task, the dataset is divided into an 80% training set and a 20% test set. The goal is to predict the time until a patient experiences a relapse or death, which is a critical aspect of treatment planning and prognosis.

The model's performance in this survival task is assessed using survival analysis techniques, including the Kaplan-Meier estimator and concordance index (C-index). These metrics help determine how well the model can predict the LFS, thus providing insights into its potential clinical applicability.

The third experiment involves unsupervised clustering applied to the entire dataset. This task is designed to assess the model's ability to discover inherent patterns and groupings within the data without prior labels. The unsupervised clustering results are evaluated based on cluster purity, silhouette score, and other relevant clustering metrics.

This experiment aims to uncover underlying biological insights and identify potential new subtypes or groupings within the data that may not have been previously recognized. The clustering results are analyzed in the context of known clinical and biological characteristics, providing a deeper understanding of the data's structure.

3.2.5 . Subtypes Classification

Myelodysplastic Syndromes (MDS) and Chronic Myelomonocytic Leukemia (CMML) are distinct hematological disorders, each with unique characteristics, as classified by the 2016 World Health Or-

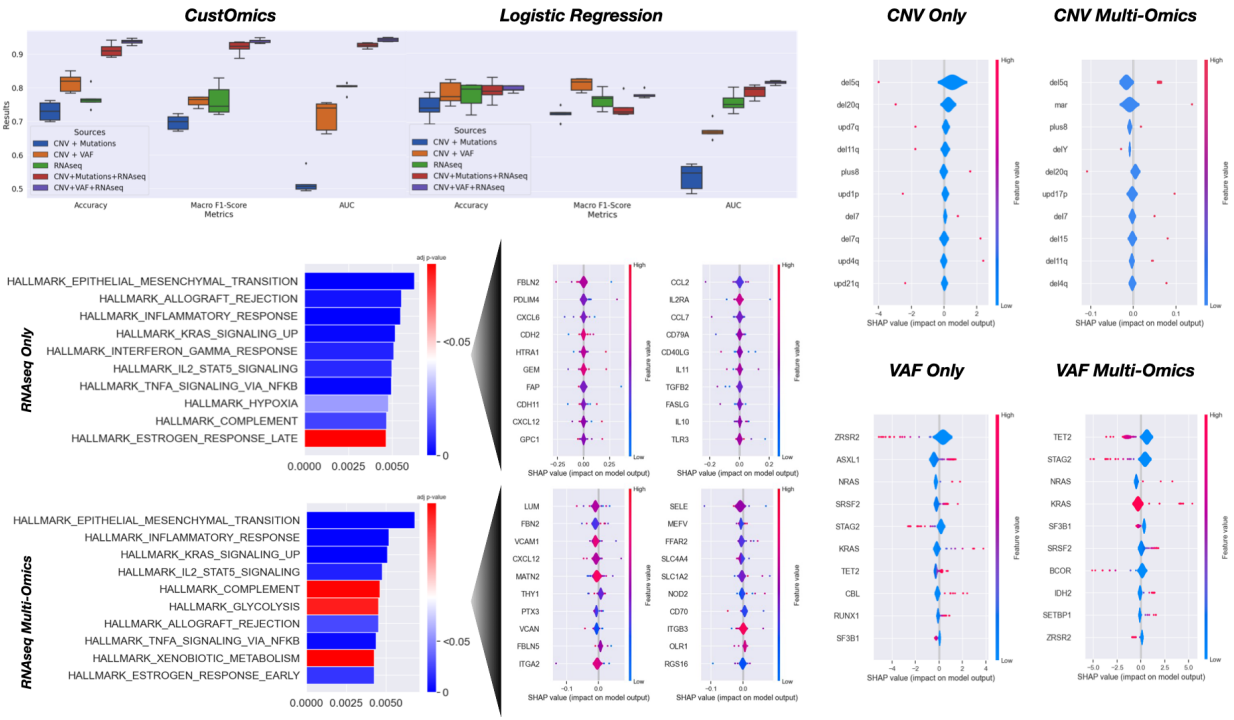


Figure 3.6: ***a.*** *Comparison Study between CustOmics and Logistic Regression for the discrimination between MDS and CMML subtypes with multi-omics data.* ***b.*** *Pathway Enrichement Analysis in single and multi-omics setup with gene importances of the top 2 pathways.* ***c.*** *Gene importance for CNV and VAF sources in single and multi-omics setup.*

ganization (WHO) criteria. MDS is primarily marked by ineffective hematopoiesis, which leads to a reduced production of blood cells and an increased risk of progression to acute myeloid leukemia (AML). On the other hand, CMML is a hybrid disorder that exhibits both myelodysplastic and myeloproliferative features. Myelodysplastic features in CMML include abnormal blood cell production and dysplasia (abnormal development of blood cells), while myeloproliferative features involve an excessive production of certain blood cells, particularly persistent monocytosis (an elevated number of monocytes in the blood). Accurately distinguishing between MDS and CMML is essential, as it guides the appropriate treatment strategies and informs the prognosis for the patient.

Figure 3.6 illustrates the performance of the CustOmics framework compared to a standard logistic regression model across various metrics, including accuracy, macro F1-score, and area under the curve (AUC). The box plots indicate that CustOmics consistently outperforms logistic regression as more modalities are integrated. For instance, while logistic regression shows marginal improvements with adding each data type, CustOmics substantially increases performance metrics, particu-

larly when combining CNV, VAF, and RNA-seq data. This suggests that CustOmics is better equipped to leverage the complex, high-dimensional data from multiple omics sources, resulting in the more accurate and robust classification of MDS and CMML.

The pathway analysis reveals significant differences between using RNA-seq data alone and integrating it with other omics data. When considering RNA-seq only, key pathways such as epithelial-mesenchymal transition, inflammatory response, and TNF signaling via NFB are highlighted. These pathways are critical in the pathophysiology of myeloid malignancies and reflect underlying biological processes driving disease progression.

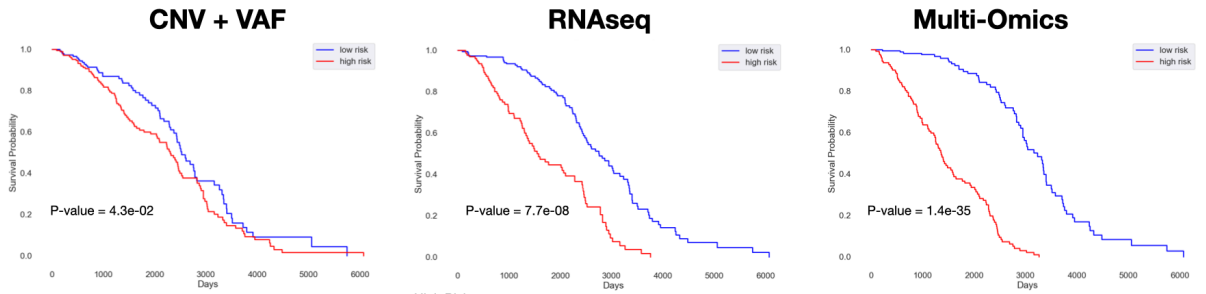
However, when RNA-seq data is combined with other omics data (multi-omics), additional pathways such as inflammatory response and IL2 STAT5 Signaling emerge as more significant than in the single-omic setup. This indicates that integrating multiple data types reinforces the understanding of known pathways and uncovers new biological insights that may be missed when analyzing RNA-seq data alone. For example, identifying glycolysis pathways highlights metabolic alterations that could be pivotal in disease mechanisms and therapeutic targeting [26, 182].

The SHAP value plots for CNV and VAF highlight the most important genes contributing to the classification model. For CNV data, critical chromosomal abnormalities such as deletions and duplications (e.g., del5q, del20q, upd7q) are emphasized. These genetic alterations are well-documented as significant factors in MDS and CMML pathogenesis. For example, del5q is associated with a distinct MDS subtype with specific clinical and therapeutic implications [26].

For VAF data, key mutations in genes such as *TET2*, *ASXL1*, and *SF3B1* are highlighted. These genes are commonly mutated in myeloid malignancies and play crucial roles in epigenetic regulation, splicing, and transcriptional control. Mutations in *TET2* and *ASXL1* are particularly prevalent in CMML, where they contribute to clonal hematopoiesis and disease progression [182].

3.2.6 . Survival Outcome Prediction

We conducted survival outcome predictions using various combinations of omics data to assess the impact of multi-omics integration on risk stratification based on Leukemia-Free Progression (LFS). Figure 3.7 illustrates the Kaplan-Meier survival curves for patients classified as high-risk and low-risk based on our predictive models. The integration of multi-omics data demonstrated a marked improvement in distinguishing between high and low-risk patients compared to models using single-omics data alone. This is evidenced by the significantly lower log-rank p-value observed in the multi-



High Risk

	0	1000	2000	3000	4000	5000	6000
At risk	199	109	63	20	5	1	1
Censored	4	67	84	96	98	98	98
Events	0	27	56	87	100	104	104

Low Risk

	0	1000	2000	3000	4000	5000	6000
At risk	162	89	57	20	5	3	1
Censored	1	54	75	88	91	93	93
Events	0	20	31	55	67	67	69

High Risk

	0	1000	2000	3000	4000	5000	6000
At risk	153	71	36	9	0	0	0
Censored	3	53	64	70	71	71	71
Events	0	32	56	77	85	85	85

Low Risk

	0	1000	2000	3000	4000	5000	6000
At risk	198	124	74	31	8	3	1
Censored	2	65	98	113	118	120	120
Events	0	11	28	56	74	77	79

High Risk

	0	1000	2000	3000	4000	5000	6000
At risk	154	70	34	3	0	0	0
Censored	1	44	50	51	51	51	51
Events	0	41	71	101	104	104	104

Low Risk

	0	1000	2000	3000	4000	5000	6000
At risk	154	70	34	3	0	0	0
Censored	1	44	50	51	51	51	51
Events	0	41	71	101	104	104	104

Figure 3.7: ***Multi-Omics Kaplan Meier:*** *Kaplan-Meier survival curves comparing the impact of different omics data combinations on Leukemia Free Survival (LFS). The three panels represent the survival analysis for different sets of omics data: (1) CNV + VAF, (2) RNAseq, and (3) Multi-Omics. For each dataset, patients are stratified into high-risk and low-risk groups based on the median of the risk score generated by the model. The blue curves represent the survival probability of the low-risk group, while the red curves represent the high-risk group. The associated p-values indicate the statistical significance of the differences between the high and low-risk groups. The corresponding tables below each Kaplan-Meier plot detail the number of patients at risk, censored, and the number of events (relapse or death) over time for both risk groups.*

omics integration approach, indicating a more robust and statistically significant separation of survival curves.

To further explore this difference in survival, we analyze the explainability results of this study. Figure 3.8 shows that the hallmark pathway enrichment analysis from RNA sequencing data, both alone and combined with other omics, highlights significant pathways such as *EPITHELIAL MESENCHYMAL TRANSITION*, *ALLOGRAFT REJECTION*, and *KRAS SIGNALING UP*, which have been previously implicated in cancer progression and immune response [149, 230]. These pathways are crucial in influencing LFS, indicating that cell migration, immune system modulation, and oncogenic signaling play vital roles in disease progression and patient outcomes.

The SHAP (Shapley Additive exPlanations) values for copy number variations (CNVs) reveal that deletions in chromosomes 5q, 7q, and 20q are significant predictors of LFS. These CNVs have been extensively studied and are known to be associated with poor prognosis in MDS [191, 131]. Their identification underscores their critical role in the pathogenesis and prognosis of the disease, as deletions in these regions often result in the loss of tumor suppressor genes and other regulatory elements

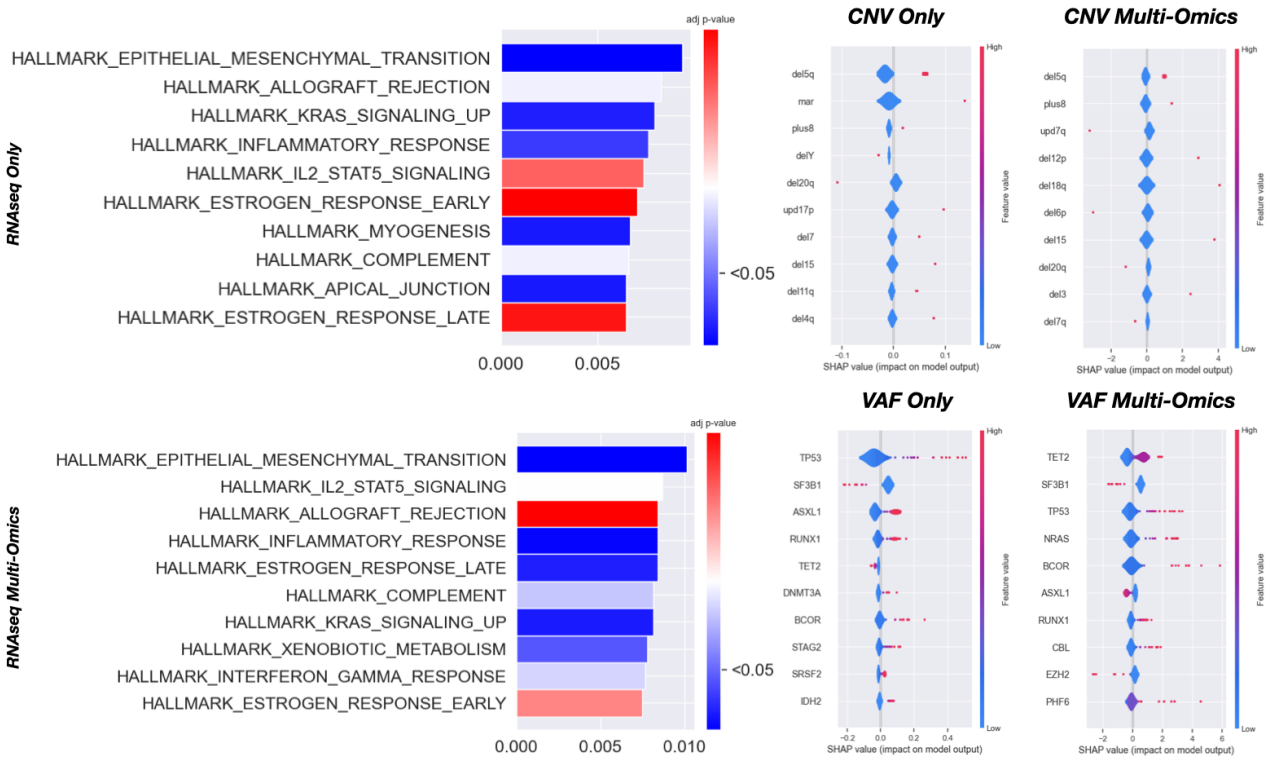


Figure 3.8: ***Multi-Omics Survival Analysis:*** *Explainability results for survival outcome prediction at the pathway and gene levels across RNAseq-only and multi-omics models. (Left) Pathway-level analysis identifies significantly enriched hallmark pathways (adjusted p-value < 0.05), such as "Epithelial Mesenchymal Transition" and "IL2/STAT5 Signaling," highlighting their importance in survival prediction. (Right) Gene-level SHAP value plots illustrate the contributions of key features to the model's predictions. Notable genetic alterations, including TP53, TET2, and SF3B1 mutations, as well as CNVs (e.g., del5q and del7q), demonstrate distinct patterns of importance. Multi-omics models exhibit superior interpretability and feature relevance compared to RNAseq-only models, reinforcing the utility of integrative approaches in survival outcome modeling.*

crucial for maintaining cellular homeostasis.

Similarly, the SHAP values for variant allele frequency (VAF) data, both alone and in a multi-omics context, identify critical mutations in genes like *TP53*, *SF3B1*, *ASXL1*, and additional significant features such as *TET2*, *NRAS*, and *PHF6* when multi-omics data is included. Mutations in these genes are well-documented as pivotal in the development and progression of MDS [103, 191]. For instance, *TP53* mutations are associated with adverse outcomes due to their role in genomic instability, while mutations in *SF3B1* and *ASXL1* influence splicing and epigenetic regulation, respectively, contributing to the complex molecular landscape of MDS.

3.2.7 . Unsupervised Exploration

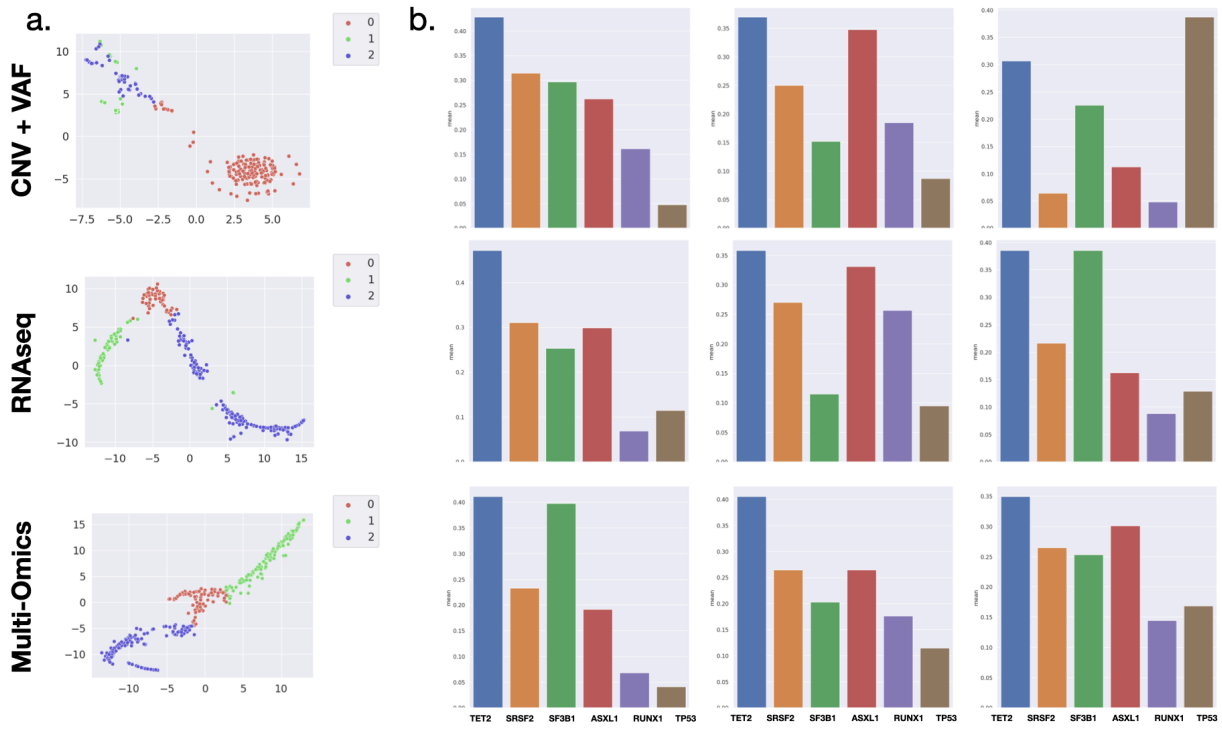


Figure 3.9: ***Multi-Omics Unsupervised Exploration:*** *We perform Unsupervised clustering on the latent space of CustOmics built in single and multi-omics setup.* ***a.*** *Latent Representation of single and multi-omics data along with overlay of the different clusters.* ***b.*** *We also display the mutation frequencies of target genes for each cluster against the rest.*

In this study, we conducted unsupervised clustering analysis on a cohort of myelodysplastic syndrome (MDS) patients using three different omics data setups: CNV + VAF, RNAseq, and Multi-Omics. The clustering results are presented in Figure 3.9a, with corresponding mutation frequency distributions shown in Figure 3.9b.

The latent space visualizations in Figure 3.9a reveal the clustering patterns for each omics setup. Three distinct clusters were identified for the CNV + VAF data, though there was some overlap between clusters 1 and 2. In contrast, the RNAseq and Multi-Omics data demonstrated better discrimination between clusters, indicating that these data types are more effective at capturing distinct biological subgroups. This enhanced clustering resolution with RNAseq and Multi-Omics likely reflects the more prosperous and detailed molecular information these approaches provide.

Figure 3.9b compares the mutation frequencies of six essential genes (TET2, SRSF2, SF3B1, ASXL1,

RUNX1, TP53) across the three identified clusters for each omics setup. For the CNV + VAF data, TET2 mutations are predominantly found in cluster 0, indicating that this mutation is a significant driver in this subgroup. Additionally, a cluster with distinctly higher frequencies of SRSF2 and SF3B1 mutations suggests a co-occurrence or synergistic role in this group, while ASXL1 and RUNX1 mutations are more common in cluster 2. The RNAseq data reveal that TET2 mutations remain most frequent in cluster 0, and similar to the CNV + VAF setup, SRSF2 and SF3B1 mutations are again highly prevalent in a distinct cluster. In the Multi-Omics data, TET2 mutations are the most frequent in cluster 0. However, in this setup, the cluster with widespread ASXL1 mutations is no longer prominent, and there is a more balanced distribution of mutations across clusters, except for SF3B1, which remains frequent in one cluster.

3.3 . Discussion

In this work, we presented a range of integration strategies for multi-source data that can handle high dimensionality and data heterogeneity. To leverage the strengths of these strategies, we introduced the mixed-integration approach and the CustOmics framework to overcome the limitations of existing methods. This new framework achieves superior latent representations, leading to a more robust and generalizable architecture, as demonstrated by the consistently better results against various integration strategies.

CustOmics adapts to each omic source by handling the training independently in the first phase, addressing the issue of unbalanced signals by standardizing the representations before learning cross-modality interactions. Our fusion model improved classification and survival outcome prediction performance across all test cases. Notably, CustOmics excelled not only on pan-cancer data but also on smaller datasets for specific cohorts, underscoring the robustness of our method in situations with fewer samples.

The application of CustOmics to integrating multi-omics data in Myelodysplastic Syndromes (MDS) highlighted the enhanced signal provided by combining multiple omics layers. The integration of RNAseq and Multi-Omics data provided better discrimination between clusters than single-omics data alone, reflecting the richer molecular information provided by these approaches.

CustOmics significantly outperformed traditional methods and other deep learning strategies in classification tasks. When distinguishing between MDS and CMML subtypes, CustOmics consis-

tently showed superior performance metrics, demonstrating its capability to leverage complex, high-dimensional data from multiple omics sources. In survival outcome predictions, integrating multi-omics data significantly improved the stratification of high-risk and low-risk patients, as evidenced by the lower log-rank p-values and improved separation of survival curves. This enhanced stratification can lead to more personalized treatment plans and better patient outcomes in clinical settings.

By adapting the SHAP method to our architecture, we highlighted essential genes for specific tasks, providing valuable biological insights. However, the computational cost of this interpretability method remains high, suggesting the need for further optimization.

Our findings suggest several future directions to enhance CustOmics. Incorporating prior knowledge into the intermediate autoencoders, such as introducing a negative binomial prior in the RNA-Seq autoencoder, could improve performance. Studying the benefits of per-source transfer learning during phase 1 could help pick up weaker signals and reduce noise. Further exploring the interpretability component of CustOmics could make the framework more actionable for clinical use. Including larger and more diverse cohorts in future studies could validate the generalizability of CustOmics across different populations and cancer types, and incorporating additional omics layers could provide a more comprehensive molecular profile.

In conclusion, our generic and interpretable multi-source deep learning framework, CustOmics, improves state-of-the-art integration strategies by proposing a hybrid approach that fits well with multi-omics data. The framework is available on GitHub: https://github.com/HakimBenkirane/CustOmics.

Analysis of Histopathology Slides

Contents

Abstract

Histopathology slides have long been the cornerstone of pathology diagnostic, providing essential insights into disease states' cellular and morphological characteristics, particularly in oncology. The digitization of these slides into whole-slide images (WSIs) marks a significant technological advancement, enabling more detailed and comprehensive analysis using computational tools. However, despite these advancements, the field faces substantial challenges, particularly in efficiently processing the vast data contained in WSIs and extracting clinically relevant information that can aid in precision medicine.

This chapter introduces three significant contributions to the field of histopathology slide analysis. First, it presents Hyper-AdaC, a novel hypergraph-based representation for whole-slide images (WSIs) that enhances traditional image analysis by modeling complex spatial and morphological relationships within tissues, thereby improving survival models' predictive power and clinical relevance. Second, the chapter introduces H&Explainer, a new tool designed to make deep learning models more transparent and interpretable, facilitating their integration into clinical workflows and improving the explainability of hypergraph-based models. Lastly, it proposes a new method for counterfactual analysis using explainable features, offering a better understanding and validation of model decisions, which is crucial for advancing precision medicine. Together, these contributions address the challenges of processing and extracting clinically relevant information from the vast data in WSIs, paving the way for more effective precision oncology.

4.1 . Related Work & Challenges

Computational pathology has undergone significant advancements in the last decade, propelled by the introduction of whole-slide image (WSI) scanners. These technologies have transformed traditional histopathology slides into high-resolution digital images. They facilitate their use in cancer diagnosis and prognosis through advanced gigapixel image analysis and statistical learning techniques [271, 150]. Notably, WSIs have been increasingly applied to various predictive tasks within oncology, with survival prediction emerging as a particularly complex challenge [282, 283]. Survival prediction models aim to estimate the time until an event, such as death or relapse, occurs, requiring a nuanced understanding of the disease processes depicted in WSIs.

The literature reveals diverse methodologies addressing the challenge of processing these large-scale images to develop robust survival models. Among these, Multiple Instance Learning (MIL) has gained prominence. MIL is a variant of supervised learning where the model is trained on sets, or "bags," of instances, with labels assigned only to the entire bag rather than individual instances. This approach is convenient in scenarios where precise instance-level annotations are unavailable or impractical, such as in medical imaging, where a whole slide image may contain malignant and benign regions. However, only a slide-level label (e.g., cancerous or non-cancerous) is provided. By leveraging the inherent variability within the instances of a bag, MIL can learn to identify patterns that contribute to the overall label, making it a powerful tool for developing survival models based on large-scale image data. (More details on the MIL framework can be found in Appendix A.1)

MIL employs a weakly-supervised learning framework where small patches from WSIs are treated as independent instances within larger, unordered groups known as bags [231, 176, 157, 260]. While MIL has shown effectiveness in tasks like cancer grading [280] and subtyping [8], its application to survival prediction is complex. Traditional MIL models often overlook the crucial integration of local and global features from WSIs, treating each instance or bag independently, which limits their ability to capture comprehensive contextual information about tumor characteristics and the surrounding microenvironment that are vital for assessing patient mortality risk [212].

To address these limitations, there has been an increased interest in employing graph-based representations (More details on Graphs can be found in Appendix A.2). These methods model the relationships between image patches using networks that facilitate interactions, potentially capturing broader tumor characteristics [2, 147, 49]. However, the practical application of graph neural networks (GNNs) in this domain faces significant challenges. The scalability of GNNs is often restricted by the computational demands of processing large graphs, which can affect their ability to generalize across different datasets [262]. Moreover, the sampling technique used to manage these large graphs can exclude critical pathological information by covering only a subset of the WSI [56, 67]. Additionally, the inherent limitation of graph structures in representing only pairwise relationships can lead to insufficient modeling of local structures, especially when there is considerable variability among the instances within a slide [86].

Recent studies have proposed several innovative methods to overcome these limitations. A method using Variational Graph Auto-encoder (VGAE) for WSI representation learning involves sampling patches,

constructing a fully connected graph from these patches, and then training this graph using GNNs. This approach efficiently uses memory and effectively captures the complex relationships within WSIs, leading to robust classification performance [69]. In recent studies, Graph Convolutional Networks (GCNs) have been extensively used to model the spatial relationships among image patches. For instance, GCNs have been applied to fully connected graphs created from WSI patches to capture the intricate interactions between different regions, enhancing the overall representation of the WSI [226]. Furthermore, heterogeneous GNNs, designed to handle different types of nodes and edges, are suitable for modeling histopathology data's diverse and complex nature. These models improve the ability to distinguish between various tissue types and cancer subtypes by leveraging the heterogeneous information present in WSIs [34]. Graph-based MIL models have been proposed to overcome the limitations of traditional MIL by incorporating the relationships between instances (patches) in a more structured way. This approach leads to better performance in tasks such as cancer detection and subtype classification [69]. Deep graph convolutional layers, including spectral and spatial methods, have been employed to process WSI graphs. These layers transform the feature space of each node (patch) and pool them into a final vector representation, which is then used for downstream tasks like classification and survival prediction [226].

Several methods have refined the MIL approach by employing clustering algorithms like K-Means to group patches before sampling to further enhance the robustness and data coverage. This strategy helps to identify distinct morphological phenotypes within WSIs and reduces dimensionality, thereby enhancing model robustness and data coverage [283, 260]. Additionally, recent studies have begun to explore correlations between small instances and the broader contexts within gigapixel images, challenging the initial assumptions of MIL [49, 221]. These innovations include hypergraph representations that extend beyond pairwise interactions, providing a more nuanced and comprehensive framework for capturing complex morphological and spatial features [66, 67].

The analysis of gigapixel whole-slide images (WSIs) presents significant challenges, particularly in the context of survival prediction. The immense size and complexity of WSIs make it challenging to process and extract clinically relevant information efficiently, often leading to computational limitations and loss of essential details. Traditional methods, such as random patch sampling, need help with these challenges, frequently resulting in suboptimal performance due to the random exclusion of potentially critical image regions. Additionally, conventional graph-based approaches are often

constrained by the limitations imposed by local structures and the sheer scale of the data.

In response to these challenges, we have developed Hyper-AdaC, a novel hypergraph-based representation designed to improve survival prediction from WSIs. Hyper-AdaC addresses these issues through three key innovations. First, it utilizes hierarchical clustering based on morphological similarity and spatial proximity to effectively summarize the information contained in WSIs, overcoming the limitations of graph size while avoiding restrictive assumptions such as a fixed number of clusters. This approach is also an efficient alternative to random patch sampling, selectively filtering the most relevant patches and reducing information loss.

Second, Hyper-AdaC leverages hypergraph representations to capture the complex local structures within WSIs more effectively than traditional graph-based methods by incorporating morphological and spatial features of the clustered instances. Finally, the method generates high-resolution attention maps that adapt to the tissue's morphology through agglomerative clustering, providing deeper insights into specific elements of the WSI, such as immunological responses, and linking these directly to survival outcomes.

4.2 . Hyper-AdaC: Adaptive clustering-based hypergraph representation of whole slide images for survival analysis

Within the scope of this study, we design, implement, and evaluate a hypergraph-based survival network for survival outcome prediction. For $1 \leq i \leq N$, let us denote by W_i, the WSI of a patient, T_i its event time, and C_i its censoring status. The goal of this study is to build and train a survival neural network $\mathcal{S}$ and to determine a function ϕ that maps the WSI into a hypergraph representation, such that $\mathcal{S}(\phi(W_i), \mathbf{\Theta}) = r_i$, with $\mathbf{\Theta}$ a set of trainable parameters and r_i the hazard rate of the time-to-event outcome of interest.

4.2.1 . Method

Hypergraph Construction

We denote by G_i a hypergraph representation of W_i such that $\phi(W_i) = G_i$. Before constructing the hypergraph, we performed automatic tissue and background separation using Otsu Binarization [157]. We then extract non-overlapping 256×256 patches x_j at $20\times$ magnification that are fed to a ResNet-18 trained using the same contrastive learning strategy (SimCLR) as in [55] that represents a

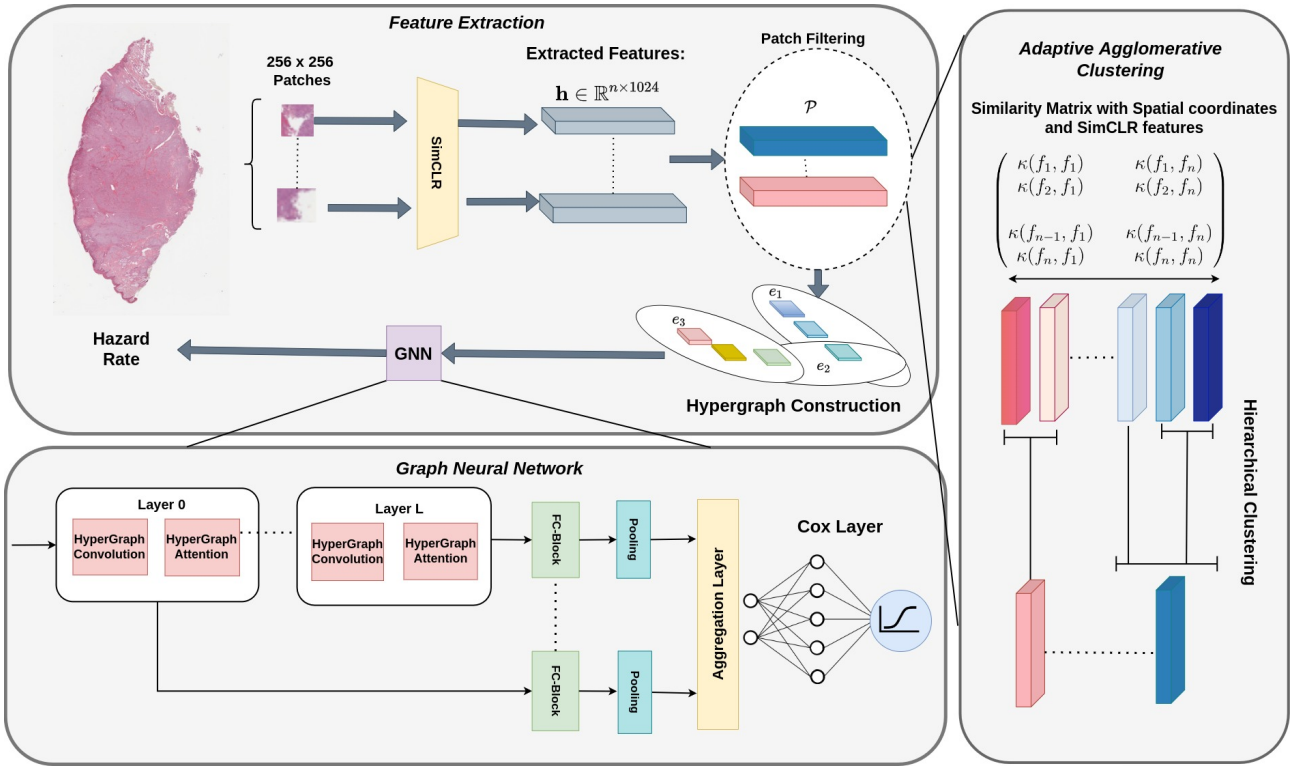


Figure 4.1: *Overview of the Hyper-adaC model*

1024-dimensional feature vector $h \in \mathbb{R}^{1024}$ each patch. The set of $(h_j)_{1 \leq j \leq n}$ associated to a W_i with n patches will be stacked into a feature matrix $\mathbf{X}_i \in \mathbb{R}^{n \times 1024}$. Each patch x_j is characterized by its ResNet-18 feature representation h_j that embeds the morphological properties of the patch and a set of coordinates $g_j = (g_{x,j}, g_{y,j})$ that represents the spatial position of the center of the patch.

Since the hypergraph should not be too large for the generalizability of the GNN [262], we perform the first step of Adaptive Agglomerative Clustering on the different patches. For that, we compute two similarity matrices $K_h \in \mathbb{R}^{n_p \times n_p}$ and $K_g \in \mathbb{R}^{n_p \times n_p}$ such that $K_h = (\kappa_h(x_i, x_j))_{1 \leq i,j \leq n_p}$ and $K_g = (\kappa_g(x_i, x_j))_{1 \leq i,j \leq n_p}$ where $\kappa_h(x_i, x_j) = e^{-\lambda_h ||h_i - h_j||^2}$ is a morphological similarity metric and $\kappa_g(x_i, x_j) = e^{-\lambda_g ||g_i - g_j||^2}$ is a spatial proximity metric.

Following the ideas presented in [159], we use the kernel $\kappa(x_i, x_j) = \kappa_h(h_i, h_j)\kappa_g(g_i, g_j)$ as a similarity kernel for agglomerative clustering. This kernel will be computed for each pair of patches from the same WSI. All patches for which similarity will be greater than a threshold δ will be considered to belong to the same cluster C_k and merged hierarchically into a single patch representation $p_k = (\tilde{h}_k, \tilde{g}_k)$ where $\tilde{h}_k = \frac{1}{|C_k|} \sum_{j \in C_k} h_j$ and $\tilde{g}_k = \frac{1}{|C_k|} \sum_{j \in C_k} g_j$.

Now that we have a reduced set of points $\mathcal{P}_i$, a hypergraph denoted by $G_i =< V_i, E_i, \mathbf{X}_i >$ is constructed. For a single WSI, we consider each clustered patch as a vertex of the hypergraph such that $V_i = [p_j]_{j \in \mathcal{P}_i}$. Each hyperedge is associated with the neighborhood of each node V_i. This neighborhood is defined as $\gamma(p_j) = \{p_k \in \mathcal{P}_i; \kappa_h(p_k, p_j) \geq \delta_h\}$, where δ_h is a threshold value to fine-tune. Those hyperedges are indicated by an incidence matrix $\mathbf{H} \in \mathbb{R}^{|\mathcal{P}_i| \times |E_i|}$ such that,

$$h(k,j) = \begin{cases} 1 & \text{if } p_j \in \gamma(p_k) \\ 0 & \text{else} \end{cases} \tag{4.1}$$

Compared to a regular graph, the exciting aspect of a hypergraph is that each node's neighborhood is depicted as a single hyperedge. This allows us to train our model with fewer parameters and thus decrease the time complexity of the convolution. In addition, it creates a community effect that gives more importance to bigger hyperedges, representing denser regions of our WSI.

Construction of the Graph Neural Network

Hypergraph Convolutions: In our proposed GNN, we employ hypergraph convolution operations [16] to capture the complex relationships within whole-slide images (WSIs). Unlike traditional graph convolutions, which operate on pairwise relationships between nodes, hypergraph convolutions extend this concept to multi-way relationships, allowing for the modeling of interactions among groups of nodes (i.e., hyperedges) rather than just pairs.

Formally, let $G = (V, E)$ represent a hypergraph, where V is the set of nodes and E is the set of hyperedges. Each hyperedge $e \in E$ connects a subset of nodes $e \subseteq V$. The incidence matrix $\mathbf{H} \in \mathbb{R}^{|V| \times |E|}$ represents the hypergraph, where $h_{ve} = 1$ if node v is connected to hyperedge e, and $h_{ve} = 0$ otherwise.

Given a node feature matrix $\mathbf{X} \in \mathbb{R}^{|V| \times d}$, where each row corresponds to a node's feature vector, the hypergraph convolution operation can be defined as:

$$\mathbf{X}' = \sigma\left(\mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{W}_e\mathbf{H}^T\mathbf{D}_e^{-1/2}\mathbf{X}\mathbf{W}\right)$$

where:

- $\mathbf{D}_v \in \mathbb{R}^{|V| \times |V|}$ is the diagonal node degree matrix with $d_v = \sum_{e \in E} h_{ve}$,
- $\mathbf{D}_e \in \mathbb{R}^{|E| \times |E|}$ is the diagonal hyperedge degree matrix with $d_e = \sum_{v \in V} h_{ve}$,

- $\mathbf{W}_e \in \mathbb{R}^{|E| \times |E|}$ is a learnable weight matrix associated with hyperedges,
- $\mathbf{W} \in \mathbb{R}^{d \times d'}$ is a learnable weight matrix associated with node features, and
- σ is a non-linear activation function, such as ReLU.

This operation aggregates features across the hyperedges, enabling the network to learn high-level representations that capture the multi-way relationships present in the hypergraph.

Hypergraph Attention Mechanism: To enhance the expressiveness of the hypergraph convolution, we incorporate a hypergraph attention mechanism [16] that allows the network to learn the relative importance of different nodes within a hyperedge. This attention mechanism assigns varying weights to the contributions of neighboring nodes, enabling the model to focus on the most relevant features during the convolution process.

Given the feature vector $\mathbf{x}_i$ of node i and the feature vectors $\mathbf{x}_j$ of its neighboring nodes j within a hyperedge e, the attention coefficient α_{ij}^e between node i and j within hyperedge e is computed as follows:

$$\alpha_{ij}^e = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T \left[\mathbf{W}\mathbf{x}_i \| \mathbf{W}\mathbf{x}_j\right]\right)\right)}{\sum_{k \in \mathcal{N}(i,e)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^T \left[\mathbf{W}\mathbf{x}_i \| \mathbf{W}\mathbf{x}_k\right]\right)\right)}$$

where:

- $\mathbf{a} \in \mathbb{R}^{2d'}$ is a learnable attention vector,
- $\|$ denotes the concatenation operation,
- $\mathbf{W} \in \mathbb{R}^{d \times d'}$ is a learnable weight matrix shared across nodes,
- $\mathcal{N}(i, e)$ denotes the set of neighboring nodes of i within hyperedge e, and
- LeakyReLU is the activation function applied to the linear transformation of the concatenated node features.

The attention coefficients α_{ij}^e are then used to compute the updated feature representation for node i as follows:

$$\mathbf{x}_i' = \sigma\left(\sum_{e \in E_i} \sum_{j \in \mathcal{N}(i,e)} \alpha_{ij}^e \mathbf{W}\mathbf{x}_j\right)$$

where E_i denotes the set of hyperedges containing node i.

This attention mechanism allows the model to dynamically adjust the influence of different nodes based on their relevance to the task, leading to more accurate and interpretable feature representations.

Final Node Representation and Pooling: Each layer consists of batch normalization and dropout layers to avoid instability during training. We also use the idea introduced in [159] of accumulating the feature representations of the convolution layers in the GNN. Those node-level representations are then pooled to generate a graph-level representation. This representation is then fed to a survival network composed of multi-layer perceptrons that predict the hazard rate used for survival outcome prediction.

Network's Loss Function: The entire network is trained using the Cox proportional hazard loss introduced in [53]; it uses the partial log-likelihood as the cost function, defined as follows:

$$\begin{aligned} pl(\mathbf{\Theta}) = \frac{1}{|\{i : C_i = 1\}|} \sum_{i:C_i=1} [\mathcal{S}(\phi(W_i), \mathbf{\Theta})) \\ -log \sum_{T_i \geq T_j} exp(\mathcal{S}(\phi(W_j), \mathbf{\Theta}))] \end{aligned} \tag{4.2}$$

where ϕ is a neural network modelling the hazard ratio and $\mathbf{\Theta}$ are the network's parameters. The cost function to train the model, to which we use a L2 regularization, is therefore defined by:

$$\mathcal{L}(\mathbf{\Theta}) = pl(\mathbf{\Theta}) + \lambda ||\mathbf{\Theta}||_2^2 \tag{4.3}$$

4.2.2 . Experimental Setup

Table 4.1: *A detailed description of the cohorts used for the study. The table includes the different cancer types, as well as the number of patients and WSIs per type.*

Cancer Type	# of Patients	# of WSIs
Bladder Urothelial Carcinoma (BLCA)	437	457
Breast Invasive Carcinoma (BRCA)	1022	1133
Glioblastoma & Lower Grade Glioma (GBMLGG)	1011	1704
Lung Adenocarcinoma (LUAD)	515	541
Ulterine Corpus Endometrial Carcinoma (UCEC)	538	566

Dataset Description

For this study, we performed extensive experiments using five different cohorts from The Cancer Genome Atlas (TCGA) detailed in Table 4.1. We chose those five datasets based on size and censoring rate. On average, each WSI contains approximately 12691 patches at $20\times$ magnification that are then reduced by hierarchical clustering to around 3147 points.

Implementation Details

The architecture of the GNN is constructed using three hypergraph convolution layers of 256 neurons, each followed by a three layers survival network of respectively 256, 128, and 64 neurons with ReLU activation that outputs the hazard ratio using a sigmoid activation function in the output layer. The entire architecture is built using fully-connected blocks. For each layer, we use a batch normalization layer to address the problem of internal covariate shift. Also, to avoid overfitting problems, we use dropout with a rate of 0.2.

For the graph construction, we select a similarity threshold of 80% with $\lambda_h = 3\lambda_g$ to give more importance to morphological features during the clustering. This choice of hyperparameters has been validated with the experiments presented in Appendix A. To train Hyper-AdaC, we used Adam optimization with a learning rate of 10^{-3} with an exponential scheduler, a weight decay of 10^{-5}, and 20 epochs.

All models were trained using an Nvidia Tesla V100S with 32 GB of memory.

Evaluation

To evaluate Hyper-AdaC, we perform 5-fold cross-validation for each cancer type. We compute the concordance index (C-index) [241] across all the validation folds to measure the predictive performance of the method. We also compare our proposed method to other state-of-the-art methods

for the same task. For all our experiments and a fair comparison, we used the same survival loss function, the exact SimCLR feature embeddings, and training hyperparameters for all methods. The benchmark methods we consider are the following::

- **DeepAttnMISL** [260]: Performs standard Multiple-Instance Learning by applying the K-Means algorithm to cluster instance-level features and then processing each cluster using Siamese networks.
- **DeepGraphSurv** [147]: A graph-based representation over sampled patches, which uses spectral graph convolution [54] to consider the topological relationships between them. We also integrate K-Means before sampling in another setup, presented as C.DeepGraphSurv on the result session.
- **Patch-GCN** [49]: Current state-of-the-art for GNN for the survival task. It performs graph multiple instance learning by considering the WSI as a 2D-point cloud, building a k-nearest neighbors graph.
- **knn-hypergraph** [66]: k-nearest neighbors hypergraph construction using sampling of patches. We use the same pipeline as Hyper-AdaC.

4.2.3 . Results

When comparing our approach to other methods, we note that Hyper-AdaC outperforms most of these in terms of C-index (Table 4.2 and Figure 4.2). Our approach generally outperforms by at least 1.6% the overall C-index on all datasets and, more specifically, in most individual datasets (except for BLCA and GBMLGG). When comparing these results with DeepGraphSurv's results, we can immediately identify the limitations of sampling patches from WSIs, as this method is the weakest in these comparisons. It only covers around 20% of the WSI and fails to train GNNs due to significant discrepancies between sampled patches. We also witnessed a clear improvement by adding context information, as almost all the graph representations outperformed the multiple-instance learning method DeepAttnMISL.

Apart from the superior performance, our method reports better robustness, highlighted by the standard deviation between the C-index values across the five folds. One can observe that Hyper-AdaC reports the lowest standard deviation, suggesting a more robust model due to the compact

Table 4.2: *Survival prediction of state-of-the-art methods using the concordance index (C-index) on 5 TCGA cohorts: Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Glioblastoma & Lower Grade Glioma (GBMLGG), Lung Adenocarcinoma (LUAD) and Ulterine Corpus Endometrial Carcinoma (UCEC).*

Model	BLCA	BRCA	GBMLGG	LUAD	UCEC
DeepAttnMISL [260]	0.514 ± 0.052	0.564 ± 0.050	0.781 ± 0.037	0.558 ± 0.060	0.595 ± 0.067
DeepGraphSurv [147]	0.495 ± 0.045	0.551 ± 0.077	0.816 ± 0.031	0.563 ± 0.050	0.614 ± 0.052
C.DeepGraphSurv [147]	0.504 ± 0.042	0.564 ± 0.043	0.787 ± 0.028	0.559 ± 0.036	0.625 ± 0.057
Patch-GCN [49]	0.561 ± 0.042	0.587 ± 0.043	$\mathbf{0.834 \pm 0.029}$	0.570 ± 0.050	0.632 ± 0.059
k-nn Hypergraph [66]	$\mathbf{0.611 \pm 0.049}$	0.545 ± 0.071	0.805 ± 0.044	0.584 ± 0.061	0.615 ± 0.020
Hyper-AdaC (ours)	0.564 ± 0.034	$\mathbf{0.592 \pm 0.025}$	0.778 ± 0.024	$\mathbf{0.595 \pm 0.012}$	$\mathbf{0.667 \pm 0.022}$

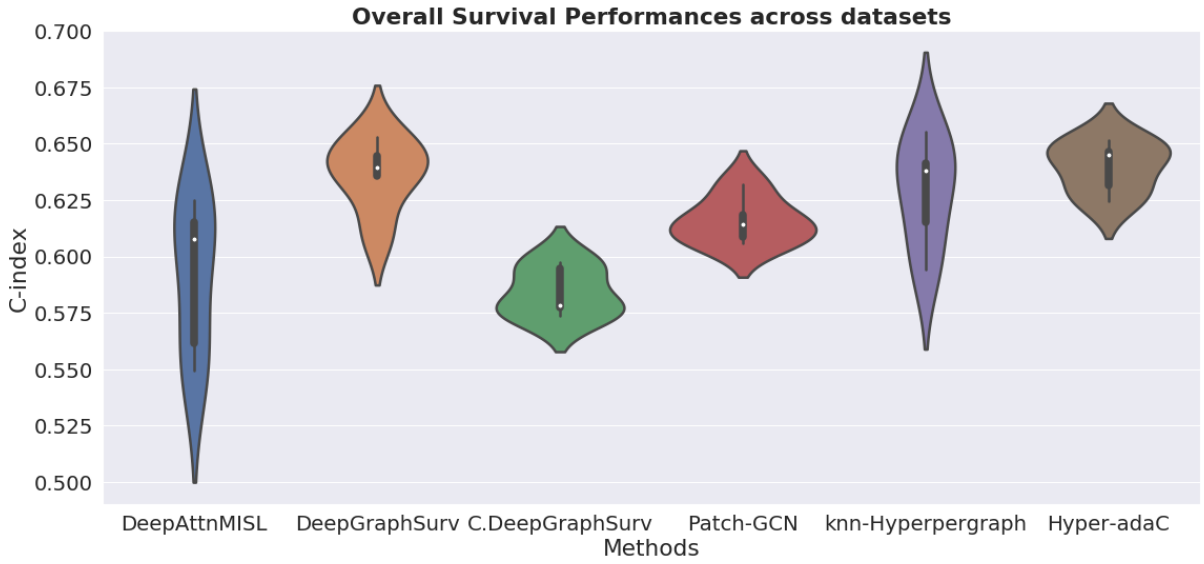


Figure 4.2: *Survival prediction performances across all datasets. They are computed by taking the C-indices on all folds of the evaluation, for all datasets.*

form of its representation. Moreover, as the representation is more miniature on Hyper-AdaC, the computing time is lower than considering the entire WSI graph since the graph convolution has the worst-case complexity of $O(n^3)$ where n is the number of nodes. However, this reduction comes with a trade-off since the graph construction part is heavier due to the hierarchical clustering step that comes with the additional complexity of $\mathcal{O}(kn^2)$, where k is the final number of clusters and n is the initial number of patches. In practice, our method is about 30% slower than graphs constructed using the whole WSI like Patch-GCN or random sampling like DeepGraphSurv. On the other hand, we are almost 20% faster during training due to more compact representations and fewer parameters.

Finally, when we compare the adaptive clustering to k-means through C.DeepGraphSurv, we observe that the adaptive property of the hierarchical clustering compared to K-means provides us with more information as it sums up quite well the discrepancies in the tissue without having to include the number of clusters as it can be adapted one slide to another, depending on the texture.

Our experiments indicate lower performances on BLCA and GBMLGG datasets. We performed additional experiments detailed in Appendix C.3.1 to analyze this point more. In fact, for the BLCA dataset, the number of elements retained after agglomerative clustering remains too high, leading to a larger and more complex graph. This increased graph size ultimately results in weaker performance. This reasoning can be inverted for GBMLGG, for which agglomerative clustering conserves only very little information, meaning that the morphological structure of this particular cancer is more homogeneous than others, and we lose much information as this clustering disregards local variability. To alleviate these issues, dataset-specific hyperparameter tuning can be performed (while we initially preferred common hyperparameters for all datasets to enhance the generalizability of our model). In practice, we add more constraints on the graph construction for the BLCA dataset by setting the similarity threshold δ to 85% and relax them on the GBMLGG dataset where δ was set to 70%. We also set $\lambda_h = 2\lambda_g$ for the GBMLGG dataset to focus less on morphological properties since the tissue is generally highly homogeneous and the clustering will be more uniform across the WSI. By doing this, we can witness a spike in performance as the C-index for our method in the BLCA dataset gets to 0.619 ± 0.037 and 0.812 ± 0.025 for the GBMLGG dataset, similar to the state-of-the-art results.

Examples of WSIs annotated by a pathologist and the corresponding model attention heatmaps are presented in Figure 4.6. We can observe that our model succeeds in discriminating zones based on their morphological and spatial features. Agglomerative clustering, by being able to adapt the number of clusters to the WSI, enables us to output attention maps that adapt well to the morphology of the slide, focusing on more relevant information and thus providing more precise information on critical local regions.

Moreover, the tumoral zone indicated by the pathologist in the first column of Figure 4.6 matches the regions where attention is at its highest. In addition, the model focused on dense inflammatory cell regions for patients with low predicted risk, which are signs of good immunity response. The multiple purple dots highlight those inflammatory cell regions in high-attention regions for the two low-risk patients (third column of Figure 4.6 showing a zoom of the attended patches). For high-risk

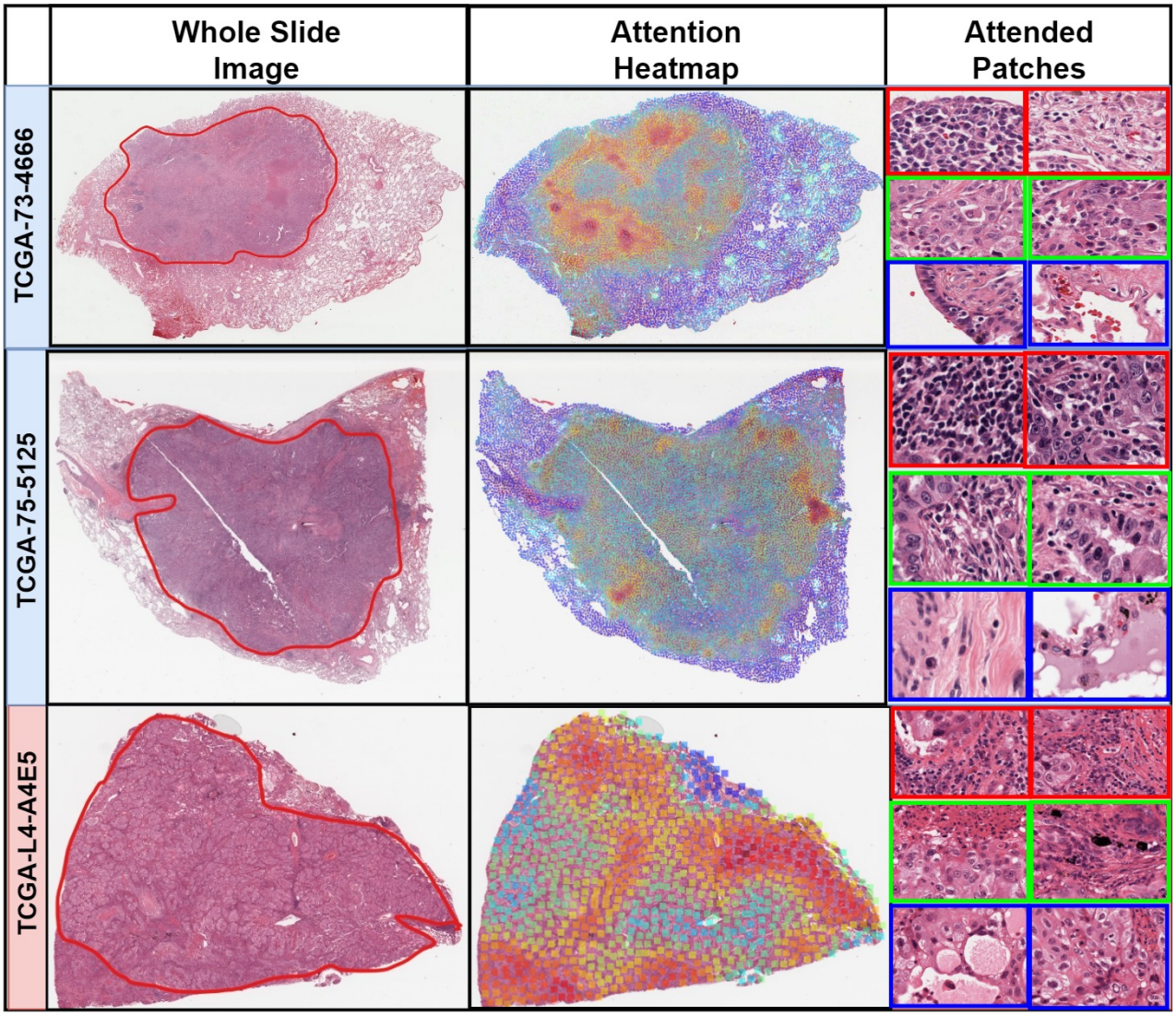


Figure 4.3: *Comparison between the model attention heatmaps and manual annotations of tumor regions for three different patients from the TCGA-LUAD dataset (blue for low-risk patients and red for high-risk patients). First column: annotations of tumor regions (in red) are superimposed in the WSI. Second column: attention heatmaps. Third column: sampled patches from 3 different attention regions; high attention (red border), medium attention (green border), and low attention (blue Border).*

patients, regions of tumor cells contain more attention due to their density. This is where the hypergraph construction presents its advantage: it creates a community behavior with hyperedges. It can assess the density of small regions through their weights. Thanks to messages passing between hyperedges, areas with more significant communities have a more decisive influence on survival pre-

diction.

However, even though our method can explain slides at high-resolution thanks to adaptive clustering, attention heatmaps can sometimes be challenging to interpret due to the high dimensionality of whole-slide images (WSIs). The complexity and vast scale of these images can obscure the clarity of the insights provided by the heatmaps, making it challenging for medical professionals to extract meaningful information. To address this issue and enhance the transparency of histopathology results, it is essential to develop methods that incorporate human-explainable features.

By leveraging human-explainable features—those that are directly interpretable by pathologists and align with established medical knowledge—we can improve the interpretability of the model's outputs. Integrating these features into the analysis of WSIs allows for a more intuitive understanding of the tissue characteristics and the model's decision-making process. This approach facilitates the transition from purely computational insights to those that clinicians can easily understand and validate, ultimately leading to more transparent and clinically applicable results in histopathology.

4.3 . Explainability Analysis on Histopathology Slides

Multiple Instance Learning (MIL) models are essential for Whole Slide Image (WSI) analysis but face significant challenges regarding interpretability. Traditional MIL methods often aggregate instance-level features for slide-level predictions without providing clear insights into which instances (patches) are most critical. This lack of fine-grained interpretability hampers their clinical utility. Conventional MIL techniques struggle with coarse attention mechanisms without detailed pathological interpretations. SI-MIL [14] attempts to address this by integrating modules that select and focus on the most relevant patches, thus enhancing interpretability by highlighting significant pathological features.

To tackle the problem of spurious associations, a counterfactual inference-based MIL framework improves interpretability by distinguishing between causal and spurious features, thereby enhancing both bag and instance-level prediction accuracy [151]. The CaMIL framework refines this approach by using causal inference techniques to block spurious associations, ensuring that the features the model learns are genuinely relevant to the pathology [48].

Attention mechanisms have notably improved the interpretability of MIL models by highlighting the most relevant patches, making the models' decisions more understandable to pathologists. Proposed by Ilse et al., attention-based MIL assigns weights to patches based on their importance, which

are then visualized to show which regions are most influential for the classification [120]. These methods have improved interpretability by clearly indicating which patches contribute most to the model's decisions.

Gradient-based explanations, as employed in the CHOWDER model, attribute features to specific patches, helping to visualize which areas are crucial for the classification [196]. This method aids in understanding the model's decision process at a granular level.

While attention-based approaches have been prominent in the literature for state-of-the-art models [221, 49], these methods generally need more detailed interpretability due to the impracticality of focusing on specific regions or cells in large images. Consequently, there is a need for more quantitative methods and tools to analyze WSIs, offering a better understanding of class differences within slides at a larger scale.

Transforming WSIs into tabular data containing human-interpretable features is crucial for enhancing the model's utility in clinical settings. As detailed in [220], this involves extracting features from densely mapped cancer pathology slides and representing them in a format that pathologists can readily understand and use for predictive tasks. This transformation makes the model's outputs more interpretable and facilitates the integration of these predictive tools into routine clinical workflows, thereby enhancing their practical applicability.

4.3.1 . H&Explainer: A Human-Interpretable Tool for Histopathology Analysis

In precision medicine, accurately analyzing Whole Slide Images (WSIs) is essential, but their size and complexity present challenges for traditional methods. While deep learning has proven effective in extracting features from these images, its "black-box" nature complicates interpretation.

To overcome this, we developed H&Explainer, a tool for explainable feature extraction from WSIs. This tool aims to clarify the deep learning models' decision-making processes, making the analysis more transparent and interpretable. By integrating multi-omics data, H&Explainer enriches the feature set, enhancing its application in oncology. This development bridges the gap between complex computational models and clinical practice, ensuring that profound learning advancements are understandable and actionable in medical contexts.

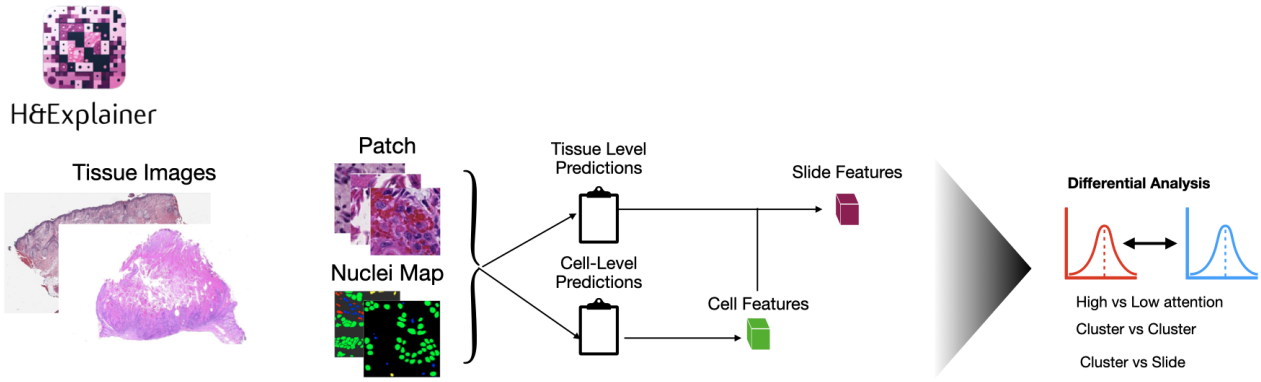


Figure 4.4: *Overview of the H&Explainer workflow*

Methods

The H&Explainer tool, as illustrated in Figure 4.4, initiates its process by performing patch extraction on whole slide histopathology images using the CellViT framework [119]. This framework employs Vision Transformers, which excel at precise cell segmentation and classification by leveraging attention mechanisms that focus on relevant parts of the image. The patch extraction process utilizes the OpenSlide library, enhanced by the RAPIDS cuCIM framework, to efficiently extract overlapping square patches (e.g., 256x256 pixels with a 64-pixel overlap) from the whole slide images. These overlaps ensure comprehensive coverage and smooth transitions between patches, reducing boundary artifacts that could compromise segmentation and classification accuracy.

For each cell identified within these patches, H&Explainer computes a range of histomic features, categorized into four principal groups: Fourier shape descriptors, Haralick features, gradient features, and morphological features [161].

1. Fourier Shape Descriptors: These features analyze the shape characteristics of cells by transforming their boundary into the frequency domain. Mathematically, if $C(t)$ represents the contour of a cell parameterized by t, the Fourier descriptors F_n are computed as the coefficients of the Fourier series expansion of $C(t)$. This allows for the compact representation of shape characteristics invariant to translation, rotation, and scaling.

$$F_n = \frac{1}{T} \int_0^T C(t) e^{-j2\pi nt/T} \, dt$$

2. Haralick Features: These texture features are derived from the Gray-Level Co-occurrence Ma-

trix (GLCM), which captures the frequency of pixel pairs with specific values and spatial relationships within the cell image. Given the GLCM $P(i, j, d, \theta)$, where i and j represent pixel intensity values at a distance d and angle θ, Haralick features such as contrast, correlation, energy, and homogeneity are computed to quantify the texture.

$$\text{Contrast} = \sum_{i,j} (i - j)^2 P(i, j)$$

3. Gradient Features: These features measure the intensity changes and edges within the cell images. The gradient $\nabla I(x, y)$ at each pixel is computed using finite difference methods, and features such as gradient magnitude and direction are derived. The gradient magnitude is given by:

$$|\nabla I(x, y)| = \sqrt{\left(\frac{\partial I}{\partial x}\right)^2 + \left(\frac{\partial I}{\partial y}\right)^2}$$

4. Morphological Features: These include basic geometrical properties such as area, perimeter, eccentricity, and solidity of the cells. For instance, the area A of a cell is computed as the number of pixels within its boundary, while the perimeter P is the total length of the boundary. Eccentricity e is computed as the ratio of the distance between the foci of the cell's boundary and its major axis length.

$$e = \frac{\text{distance between foci}}{\text{major axis length}}$$

In addition to cell-level analysis, H&Explainer provides an interactive feature for selecting Regions of Interest (ROIs) within the whole slide image. Users can specify ROIs to compute region-level features based on the aggregated histomic features of cells within these areas. This functionality is handy for analyzing specific areas of interest, such as tumor margins, stromal regions, or necrotic zones, where cellular characteristics may differ significantly from the rest of the tissue.

Moreover, the tool supports the computation of slide-level features by contrasting cell populations across predefined regions. For example, by comparing the histomic features of cells in tumoral versus stromal regions, researchers can identify patterns associated with disease progression, treatment response, or other clinically relevant outcomes. This multi-scale approach—from individual cells to entire tissue regions—enables a comprehensive analysis of the histopathology slides, providing deep

insights into the underlying biology and pathology.

Histopathology Slide Analysis

The analysis of histopathology slides in H&Explainer is multifaceted, providing detailed comparisons between different Regions of Interest (ROIs) and comprehensive assessments at the slide level. When ROIs are selected, differential analysis between cell populations in these regions can be performed by computing the differences in the distribution of each computed cell feature using a t-test. For each feature X_i, the distribution is characterized by statistical measures such as mean, standard deviation, kurtosis, skewness, and entropy. To compare the distributions of a feature X_i between two ROIs, the t-test evaluates whether the means of the two distributions are significantly different:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the means, s_1^2 and s_2^2 are the variances, and n_1 and n_2 are the sample sizes for the two ROIs. This type of analysis is particularly beneficial for interpreting attention heatmaps generated by deep learning models, as it allows researchers to assess differences in cell populations between high and low-attention regions. By understanding these differences, insights can be gained into the areas of the slide critical for the model's decision, potentially uncovering underlying biological phenomena.

Characterizing homogeneous cluster regions within the slide provides valuable information about tissue architecture and the spatial organization of different cell types. Identifying regions with similar cellular compositions can correlate with specific pathological states or responses to treatment. Clustering algorithms, such as k-means or hierarchical clustering, can be applied to a slide's feature set $\{X_i\}$ to identify these homogeneous regions, revealing patterns indicative of disease or treatment response.

At the slide level, H&Explainer facilitates the analysis of the entire slide's dynamics for predictive tasks. By performing a t-test on each slide-level feature X_i across different classes, we can identify significantly different features between a specific class and others. For a feature X_i, this involves comparing its distribution between a target class and all other classes, helping to identify distinguishing characteristics of the tissue associated with specific clinical outcomes or disease states.

$$t_{\text{class}} = \frac{\bar{X}_{\text{class}} - \bar{X}_{\text{others}}}{\sqrt{\frac{s^2_{\text{class}}}{n_{\text{class}}} + \frac{s^2_{\text{others}}}{n_{\text{others}}}}}$$

Such comprehensive analyses provide a deeper understanding of the histopathological characteristics of the tissue, aiding in diagnosis, prognosis, and treatment planning. The ability to analyze and interpret these features at both the ROI and slide levels allows H&Explainer to offer valuable insights into the cellular and structural landscape of the tissue, facilitating more informed clinical decision-making.

Experimental Setup

To demonstrate the capabilities of the H&Explainer tool, we conducted two distinct tasks designed to highlight its versatility in different analytical contexts. The first task aimed to illustrate the single-slide analysis setup. In this task, we utilized Tumor Infiltrating Lymphocytes (TIL) maps, which were predicted in the context of the study by Saltz et al. [213]. TILs are a crucial component of the tumor microenvironment and are associated with the immune response to cancer. By performing a differential analysis, we compared regions within single histopathology slides that exhibited high TIL density with those showing low TIL density. This analysis enabled us to identify significant differences in the histopathological features between these regions, showcasing the H&Explainer tool's ability to pinpoint localized variations within a single slide.

The second task was designed to showcase the cohort analysis setup, focusing on the classification and prediction of TP53 mutations. TP53 is a critical tumor suppressor gene, and its mutations are prevalent in various cancers, often associated with poor prognosis. In this setup, we aimed to provide a deeper understanding of the attention heatmaps generated by state-of-the-art deep learning models, specifically CLAM [157] and TransMIL [221]. These models are known for their efficacy in processing whole-slide images and making accurate predictions. By performing a differential analysis between high and low-attention regions identified during the TP53 classification task, we could dissect and compare the specific histopathological features that each model finds critical.

Results

Figure 4.5 presents the outcomes of H&Explainer for both single slide and cohort analyses.

In the single slide analysis, high TIL regions are distinctly characterized by a high lymphocyte population, a high tumoral population, and a high ratio between the cell's major and minor axes. These

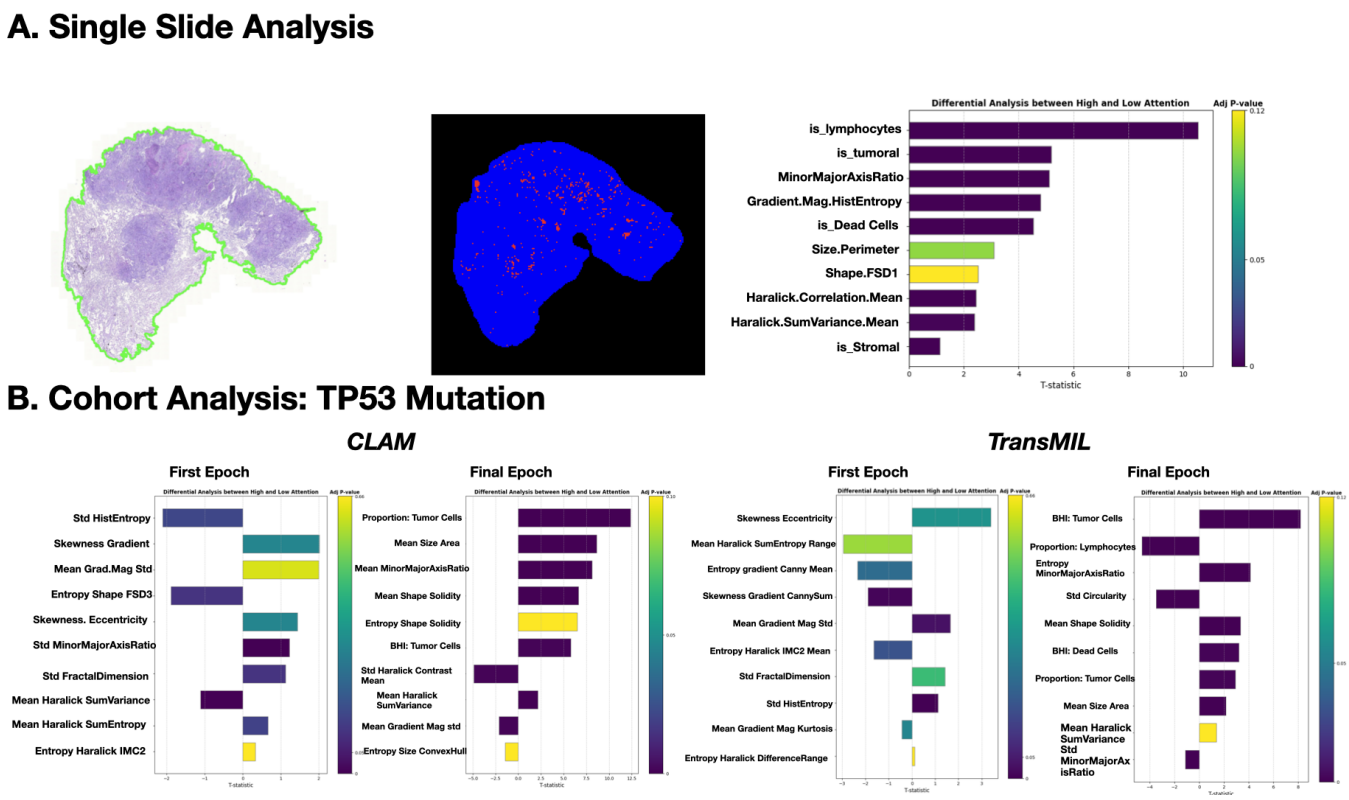


Figure 4.5: ***H&Explainer Results:*** *a. The single slide approach shows a TIL map for a histopathology slide. In this setup we compare, using a t-test, the explainable cell-level features between high and low TIL regions. b. The cohort analysis compares two models, CLAM and TransMIL for the prediction of TP53 mutations. For each model, we perform the comparison between high and low attention regions in the beginning and end of the training.*

findings highlight significant histological differences between high and low TIL regions, providing insights into the tumor microenvironment. This aligns with existing literature, demonstrating that high TIL regions are associated with an increased lymphocyte presence and tumoral activity, indicative of an active immune response against the tumor [211, 64].

In the cohort analysis, which focuses on classifying TP53 mutations using CLAM and TransMIL models, the first epoch shows no significant differences between high and low-attention regions. However, as the models train, the final epoch reveals significant differences. The CLAM model, in particular, focuses on proportional and shape attributes, such as the proportion of tumor cells and the mean area of cells, indicating its ability to capture morphological nuances associated with TP53 mutations. This observation is supported by prior research, which suggests that tumor cell morphology and size are critical indicators of genetic mutations and tumor behavior [279, 254].

In contrast, the TransMIL model emphasizes clustering properties, like the Ball-Hall Index for tumor cells, due to its approach of incorporating instance-wise correlations within a bag, unlike CLAM's independent patch instance approach. This difference enables TransMIL to capture more complex spatial relationships and clustering characteristics. Such clustering properties are significant as they correlate with tumor aggressiveness and mutation status in previous studies [3, 175]. Overall, the results demonstrate that both models improve their ability to distinguish histological features linked to TP53 mutation status with training, reflecting their adaptation to and refinement of pertinent histological attributes, thereby aligning with the training tendencies of deep learning models. These findings corroborate the growing body of evidence showing that deep Learning has the capaicty to identify and classify genetic mutations based on histopathological features [57, 74].

4.3.2 . Counterfactual Explanations For Digital Histopathology Slides Using Human Interpretable Features

Counterfactual analysis generates "what-if" scenarios by identifying minimal changes to input data that would alter a model's prediction. This technique is instrumental in enhancing the interpretability of machine learning models, particularly in high-stakes fields like healthcare, where understanding the reasoning behind model predictions is crucial.

Recent advancements in counterfactual analysis have focused on improving the efficiency and interpretability of generated explanations. For instance, MACE [258] introduces a model-agnostic framework that optimizes for minimal feature changes while ensuring plausibility. DICE [181] emphasizes diversity in counterfactuals, providing multiple plausible explanations to enhance user understanding. LACE [31] leverages latent space transformations to generate transparent and efficient counterfactuals for tabular data.

Despite these advancements, high-dimensional data remains a challenging frontier. Methods such as those by [258] and [181] often rely on greedy search algorithms, which can be prohibitively slow and may not scale well with the increasing dimensionality of data. Addressing these limitations requires novel approaches that can efficiently navigate the high-dimensional feature space while maintaining the interpretability and plausibility of the generated counterfactuals.

In response to these challenges, the use of generative models has emerged as a promising approach for counterfactual analysis, particularly in handling high-dimensional data. For example, the approach discussed in [43] utilizes generative models to construct counterfactual explanations by

modeling the distribution of the data and generating plausible counterfactuals that are close to the original input. This method demonstrates the potential of generative models to overcome the limitations of traditional greedy search techniques, providing a more scalable and efficient framework for counterfactual generation in complex, high-dimensional spaces.

In this work, we leverage human-interpretable features derived from the HExplainer tool to perform counterfactual analysis on histopathology slides. We introduce a novel method, GMM-CeFlow, which utilizes normalizing flows to map these features into a distinguishable latent space. This approach provides tractable and easily computable formulas for class transport, effectively addressing the challenge of searching for counterfactual examples in high-dimensional spaces. The integration of generative models further enhances the robustness and efficiency of our method, allowing for the generation of plausible and interpretable counterfactuals even in complex data environments.

Methods

Normalizing flows provide a powerful framework for transforming complex data distributions into simpler ones through a series of invertible mappings. In our approach, we train a mapping $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$ that transforms the original data x into a latent vector z. This mapping comprises multiple invertible functions, typically implemented using the RealNVP architecture. The likelihood of the original data $p_X(x)$ is defined using the change of variables formula:

$$p_X(x;\theta) = p_Z(f_\theta(x)) \left| \det \left(\frac{\partial f_\theta(x)}{\partial x} \right) \right|$$

Where p_Z is the density in the latent space, often chosen to be a simple distribution such as a Gaussian, the associated loss function is the negative log-likelihood:

$$-\log(p_X(x;\theta)) = -\log(p_Z(z;\theta)) - \sum_{k=1}^{K} \log \left| \det \left(\frac{\partial f_k}{\partial z_k} \right) \right|$$

In our method, GMM-CeFlow, we employ a Gaussian Mixture Model (GMM) to model the latent space, where each class corresponds to a different Gaussian distribution. This setup provides a tractable representation of decision boundaries between classes. The probability density function in the latent space conditioned on class c is given by:

$$p_Z(z|y=c) = \mathcal{N}(z|\mu_c, \Sigma_c)$$

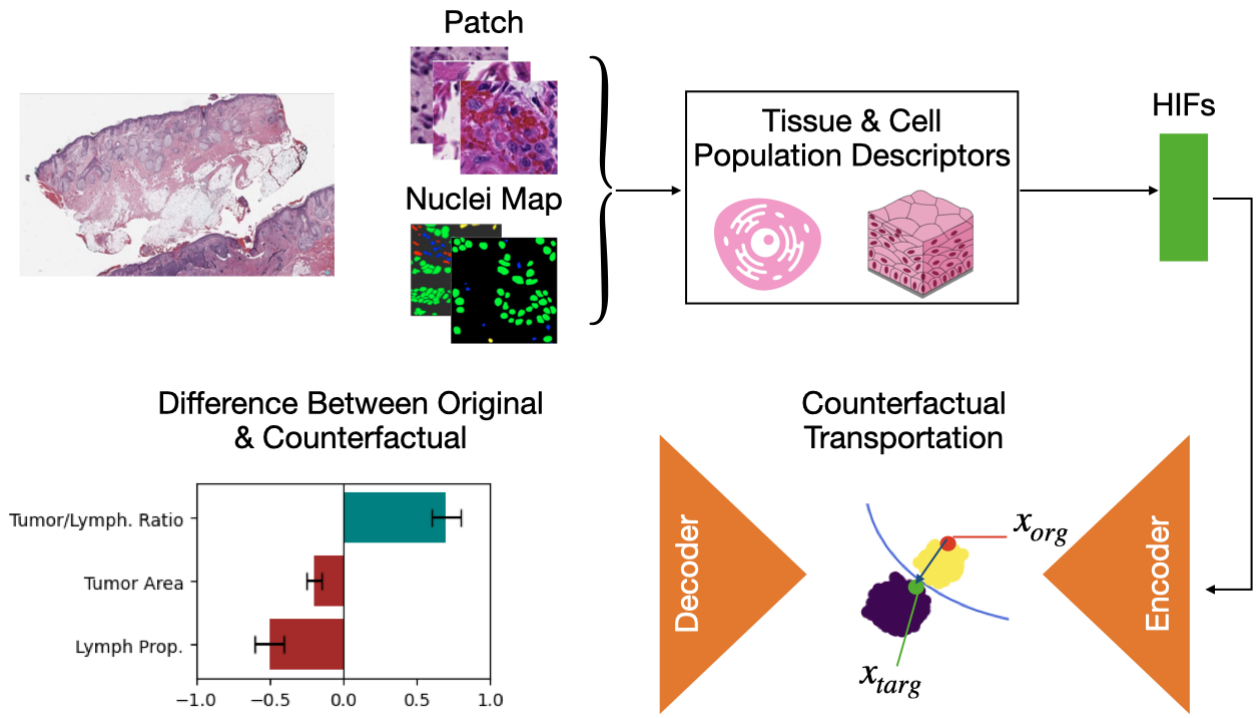


Figure 4.6: ***Overview of the Counterfactual Explanation Model's Architecture:*** *The figure illustrates a framework for counterfactual explanation in histopathology slide analysis using Human Interpretable Features (HIFs). The process starts with the extraction of patches and nuclei maps from the whole-slide image, which are subsequently used to derive Tissue and Cell Population Descriptors. These descriptors, encompassing features such as the tumor-to-lymphocyte ratio, tumor area, and lymphocyte proportion, are aggregated into Human Interpretable Features (HIFs). The HIFs feed into a counterfactual transportation framework where the original feature representation x_{org} is transformed into a counterfactual feature representation x_{targ} using an encoder-decoder architecture. The bar chart at the bottom left highlights the differences between the original and counterfactual HIFs, demonstrating how variations in these interpretable features could lead to different model predictions. This methodology offers a transparent and interpretable means of understanding how specific histopathological features influence the model's predictions, thereby enhancing the transparency and applicability of the results in clinical settings.*

For balanced classes, the overall density is defined as:

$$p_Z(z) = \frac{1}{n_c} \sum_{c=1}^{n_c} \mathcal{N}(z|\mu_c, \Sigma_c)$$

For general cases, it is:

$$p_Z(z) = \sum_{c=1}^{n_c} \pi_c \mathcal{N}(z|\mu_c, \Sigma_c)$$

where π_c are the mixture weights, and each Gaussian $\mathcal{N}(z|\mu_c, \Sigma_c)$ is isotropic, meaning $\Sigma_c = \sigma_c I_d$.

To generate counterfactual examples, we first encode the original sample into the latent space using the trained normalizing flow, $z_{org} = f_\theta(x_{org})$. The decision boundary between classes in this latent space is represented by a quadric hypersurface, defined by the equation:

$$x^T Ax + b^T x + c = 0$$

Where the parameters are given by:

$$A = -\frac{1}{2}(\Sigma_1^{-1} - \Sigma_2^{-1}), \quad b = \mu_1\Sigma_1^{-1} - \mu_2\Sigma_2^{-1}, \quad c = -\frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_2^T\Sigma_2^{-1}\mu_2) + \log\left(\frac{\alpha\pi_1}{\pi_2}\right) + \frac{1}{2}\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right)$$

The counterfactual generation process involves projecting the encoded point z_{org} onto this decision boundary. This projection ensures that the new point lies as close to the original while crossing into the desired class region. The projected point z_{targ} is then transformed back into the original data space using the inverse of the normalizing flow, yielding the counterfactual example $x_{cf} = f_\theta^{-1}(z_{targ})$.

Results

We evaluate our method on a public breast cancer dataset (TCGA-BRCA, $n = 1187$) to predict molecular subtypes using the PAM50 classifier, demonstrating the model's capability to generate multi-task counterfactual explanations. To rigorously assess the performance of our approach, we employ a set of evaluation metrics, which allows for a comprehensive comparison against other state-of-the-art methods, as outlined in [31].

The evaluation criteria encompass the following metrics:

- **Proximity between Counterfactual and Original Sample (PROX):** This metric measures how close the generated counterfactual instance is to the original sample in the feature space. A lower proximity value indicates that the counterfactual is minimally altered from the original instance, making the explanation more plausible and easier to understand in a clinical setting.
- **Average Feature Changes in Counterfactuals (CNT):** This metric quantifies the average number of features that are modified to generate a counterfactual. It provides insight into the complexity of the counterfactual, with a lower number of changes being preferable, as it suggests

that fewer modifications are required to achieve a desired outcome, thus maintaining the original instance's integrity.

- **Implausibility of Counterfactual Explanations (IMP):** IMP evaluates the biological or clinical plausibility of the generated counterfactuals. This metric checks whether the changes made to features result in realistic and meaningful explanations. High implausibility scores indicate counterfactuals that are less likely to occur in real-world scenarios, which could reduce their utility in clinical decision-making.

- **Counterfactual Success Rate in Class Change (SR):** SR assesses the effectiveness of the counterfactual in altering the predicted class. This metric reflects the percentage of counterfactuals that successfully change the classification outcome, which is crucial for the method's ability to provide actionable insights. A high success rate indicates that the counterfactuals are effective in achieving the intended outcome, thereby validating the method's utility in practical applications.

- **Average Counterfactual Generation Time (TIM):** TIM measures the computational efficiency of the counterfactual generation process, calculated as the average time taken across 10 runs. This metric is particularly important in clinical settings, where timely decision-making is critical. Faster generation times are preferable, as they enable quicker feedback and integration into clinical workflows.

Table 4.3: *Assessment of counterfactual methods using various criteria: a) Proximity between counterfactual and original sample (PROX), b) Average feature changes in counterfactuals (CNT), c) Implausibility of counterfactual explanations (IMP), d) Counterfactual success in class change (SR), e) Average counterfactual generation time across 10 runs (min) (TIM).*

Model	**PROX**	**CNT**	**IMP**	**SR**	**TIM**
MACE [258]	0.79 ± 1.00	10.54 ± 2.89	0.67 ± 0.45	91%	$103,9 \pm 7,0$
DICE [181]	0.81 ± 0.54	$\mathbf{9.27 \pm 2.51}$	0.42 ± 0.38	87%	$91,2 \pm 10,0$
T-LACE [31]	$\mathbf{0.90 \pm 0.30}$	9.31 ± 3.12	0.44 ± 0.12	97%	$82,4 \pm 9,7$
Ours	0.85 ± 0.44	9.74 ± 3.14	$\mathbf{0.39 \pm 0.22}$	$\mathbf{99\%}$	$\mathbf{14,7 \pm 3,5}$

As shown in Table 4.3.2, our method delivers performance comparable to other state-of-the-art approaches while significantly reducing computation time. This efficiency is achieved by leveraging straightforward, easily computable formulas, rather than relying on the more complex and time-consuming greedy search methods often used in other counterfactual explanation techniques.

However, it is important to highlight a limitation of our approach: we estimate the classifier's behavior using a Gaussian Mixture Model (GMM) in the latent space, which is trained alongside the classifier instead of directly using the original black-box classifier. This approach, while effective within the scope of our study, may affect the generalizability of our results. To address this concern, further validation across multiple datasets is necessary to confirm the robustness and applicability of our method in different contexts.

Moreover, while this study focuses on using interpretable histopathology features to enhance the clarity of counterfactual examples, future research should also consider exploring counterfactual analyses directly within the original image space. This could provide additional insights and further extend the applicability of our method in real-world clinical scenarios.

4.4 . Discussion

In this chapter, we introduced Hyper-adaC, a novel method employing hypergraph representation to enhance the characterization of the tumor microenvironment in histopathology slides. This innovative approach was motivated by the need to capture histopathological data's complex and high-dimensional nature more effectively. Hyper-adaC demonstrated not only promising results in terms of predictive performance but also showed robust explainability, addressing a critical gap in the current methodologies.

Despite the success of Hyper-adaC, we encountered significant challenges related to the interpretation of attention heatmaps. Histopathology slides are inherently high-dimensional, and the resulting attention heatmaps often lack clarity and interpretability. This ambiguity can hinder the practical utility of the method; as evident, actionable insights are paramount in medical decision-making.

To address this limitation, we developed two complementary methods. The first is H&Explainer, a tool designed to extract human-interpretable features and utilize them to analyze different regions of interest (ROIs) in histopathology slides. HExplainer bridges the gap between complex computational models and clinical applicability by transforming abstract attention maps into features that patholo-

gists and researchers can readily understand and evaluate. This tool significantly enhances the interpretability of hypergraph-based representations, making the insights derived from Hyper-adaC more accessible and actionable.

The second method we conceived is based on counterfactual analysis, leveraging the human-interpretable features extracted by H&Explainer. Counterfactual analysis provides a robust framework for understanding model predictions by considering alternative scenarios—asking "what if" questions. By applying this method, we can offer deeper explanations for predictive outcomes, clarifying how specific features influence the model's decisions and highlighting potential areas for clinical intervention or further investigation.

Together, these two methods enhance the interpretability and applicability of our hypergraph-based approach to analyzing histopathology slides. They allow for a more nuanced understanding of the tumor microenvironment and provide clinicians and researchers with clear, interpretable insights that can inform treatment decisions and guide further research.

In conclusion, this chapter presents significant advancements in histopathology slide analysis through the development of Hyper-AdaC, a hypergraph-based model that accurately captures the spatial and morphological complexities of the tumor microenvironment. By incorporating H&Explainer, our approach provides interpretable insights into model decisions and generates counterfactual explanations that highlight critical features driving classification outcomes. These contributions specifically address the challenges of working with high-dimensional whole-slide images, offering a more transparent approach to understanding the relationships between tissue structures and survival outcomes.

Future work will involve further optimizing Hyper-AdaC by testing it on a wider range of cancer types and datasets from sources like The Cancer Genome Atlas (TCGA) and beyond. Additionally, efforts will be made to seamlessly integrate these methods into clinical workflows, enabling real-time decision support in pathology labs. Improving the explainability of the generated insights will make the model more actionable for clinicians, ultimately contributing to more personalized treatment strategies and better patient care. This work lays a foundation for more interpretable and clinically useful applications of computational pathology.

Multimodal Integration of Multi-Omics Data & Histopathology Slides

Contents

Abstract

This chapter introduces a novel methodology, Multimodal CustOmics, designed to integrate multi-omics and histopathology data. By combining diverse data types, CustOmics provides a comprehensive framework for enhanced precision medicine in oncology. We detail the methodology's development, underlying algorithms, and integration strategies, highlighting its potential to uncover complex biological insights. A significant application of CustOmics is demonstrated using a lung cancer dataset, showcasing its effectiveness in real-world scenarios. This work underscores the transformative potential of multimodal data integration in advancing personalized cancer treatment.

5.1 . Related Work & Challenges of Multimodal Integration

In our previous work, we developed CustOmics for the integration of multi-omics data and Hyper-AdaC for modeling Whole-Slide Images (WSIs) in computational pathology. Each of these frameworks addresses specific challenges in their respective domains, with CustOmics focusing on the effective combination of various molecular data sources, and Hyper-AdaC leveraging hypergraph-based representations to model the complex spatial relationships within WSIs. Building on these advancements, our current goal is to combine these modalities into a unified framework that can seamlessly integrate multi-omics and histopathological data, enabling a more comprehensive and interpretable approach to cancer diagnosis, prognosis, and therapeutic response prediction.

In recent years, integrating whole slide images (WSIs) and omics data has garnered significant attention in computational pathology. Models that merge these diverse data types aim to leverage the complementary information from histopathological images and molecular profiles to improve diagnostic accuracy, prognostic predictions, and therapeutic decisions. Notable approaches have been proposed to harness the power of multimodal data integration. For instance, Courtiol et al. [58] proposed a method that combines deep features from histology images with transcriptomic data using a multimodal approach to enhance cancer diagnosis. Similarly, Lu et al. [158] developed a model that integrates genomic and histopathological data using a unified deep-learning architecture to predict patient outcomes in lung cancer. These studies underscore the potential of multimodal models to transform precision medicine by providing more comprehensive insights into disease mechanisms.

One of the most critical advances in the field is the work of Chen et al. [50], which presents a transformer-based model with a co-attention mechanism to merge WSIs with genomic data. This approach represents a significant leap in enhancing histopathological and genomic information integration. More recently, Jaume et al. [123] utilized the segmentation of transcriptomics data by pathways to better focus on relevant biological functions, demonstrating another innovative approach to multimodal integration.

Despite these advancements, existing methods often need help with proper multi-omics integration. Many techniques focus on single-omics integration, failing to harness the true potential of combining multiple omics data types. For example, while some studies, such as those by Porpoise [51], attempt multi-omics integration, they typically use early-integration techniques and standard feed-forward networks, which fails to exploit the rich information provided by multi-omics data fully. In chapter 3, we highlighted these limitations and advocated for more advanced integration strategies to unlock the full potential of multi-omics data.

Furthermore, a significant challenge that needs to be addressed is the issue of multimodal integration with missing modalities. In practical clinical settings, incomplete datasets are expected to be encountered where one or more modalities may need to be included. Developing robust models that can effectively handle such incomplete data is crucial for the practical application of multimodal integration techniques. Addressing these challenges will be essential for advancing the field of computational pathology and realizing the full potential of integrating WSIs and omics data for improved diagnostic and prognostic capabilities.

5.2 . Multimodal CustOmics: A Unified and Interpretable Multi-Task Deep Learning Framework for Multimodal Integrative Data Analysis in Oncology

To tackle the challenges of integrating multimodal data, we propose a deep-learning framework to create an interpretable image-omic representation, represented in 5.1, that captures interactions at multiple levels of the biological system. Named Multimodal CustOmics, this integration method builds upon the strategy introduced in chapter 3. The original network optimally integrated heterogeneous data from different omics sources while preserving the specificity of each modality. The new version of our multimodal network can jointly integrate H&E slides and molecular profile features

(mutation status, copy-number variation, RNA sequencing [RNA-seq] expression, DNA methylation, etc.). Additionally, it can interpret how the interaction between all those sources correlates with specific supervised tasks such as molecular subtype identification or survival prediction. Furthermore, this method facilitates the assessment of feature importance at multiple levels through ad-hoc scores. At the gene level, the method outputs the importance score of each gene for each molecular source, both independently and in association with other modalities. At the pathway level, a Multimodal Pathway Enrichment Score (MPES) is computed to assess the importance of a specific pathway for a specific prediction task, such as molecular subtype classification or survival prediction. This score has also been extended to account for spatial correlations in the histopathology slide and reflects the importance of the interaction between spatial regions of the WSI and pathways. We compare CustOmics to four other methods integrating omics and histopathology data. For this comparison, we follow the study conducted by Chen et al. [50] and use as a basis of comparison the engineered baselines introduced for omics, histopathology, and multi-omics integration:

- **SNN:** We train a feed-forward self-normalizing network architecture [138] as a multi-omics baseline, where we concatenate multi-omics data before feeding them to the network. This architecture, used in [50, 51], is state-of-the-art for histology-genomic integration.

- **DeepSets:** One of the first neural architectures for set-based deep-learning problems [269] proposes sum pooling over instance-level features. Its multimodal extension is presented in Chen et al. [50], where it processes omics data with an SNN and integrates them with bilinear pooling.

- **Attention MIL:** A set-based network similar to DeepSets replaces sum pooling with an attention pooling technique [120].

- **DeepAttnMISL:** A set-based network that first applies K-Means clustering to instance-level features, processes each cluster using Siamese networks, and then aggregates the cluster features using global Attention pooling [261].

- **MCAT:** In Chen et al. [50], researchers present MCAT as the current state-of-the-art in multimodal histology-genomics integration. It is a transformer-based set-based network that combines modalities with bilinear attention-based pooling.

- **SurvPATH:** Jaume et al. [123] presents a new transformer-based architecture that embeds transcriptomics information in the form of biological pathways, similarly to the method presented in this paper. The omics layer in this paper consists of an SNN architecture that we will adapt to take into account multi-omics data.

This study employed a multi-level interpretability approach focusing on gene and spatial levels. We further explain spatial interpretability results by extracting high-attention image patches and analyzing them using the pre-trained Hover-Net model for cell instance segmentation and classification [96]. This analysis categorized cells into tumors, lymphocytes, stromal, necrosis, and epithelial cells. We quantitatively assessed the frequency of these cell types in high-attention patches for each patient. Additionally, we estimated the proportion of Tumor Infiltrating Lymphocytes (TILs) in these patches using the methodology developed by Saltz et al. [213], thus providing a deeper understanding of the tumor microenvironment.

5.2.1 . Method

Within the scope of this study, we design, implement, and evaluate a multimodal integration network for integrating histopathology slides and multi-omics data. For $1 \leq i \leq N$, let us denote by W_i and O_i, respectively, the WSI and multi-omics bag for patient i. The goal of this study is to build and train a multi-task network $\mathcal{M}$ that takes as input the two bags and creates an interpretable multimodal representation z_i for each patient such that $\mathcal{M}(W_i, O_i) = z_i$.

Building upon the original Hyper-AdaC framework introduced in 4, we refine the approach to Whole Slide Image (WSI) analysis by employing a hard clustering method to delineate distinct tissue regions. In this approach, each patch is assigned to a single cluster, ensuring that the clusters represent well-defined and non-overlapping regions within the tissue. This process enables the creation of per-region embeddings that effectively capture the morphological and spatial characteristics of the tissue, while maintaining the clarity and simplicity of distinct cluster boundaries.

These hard-clustered regions are then used to construct a hypergraph, where the patches serve as nodes, and hyperedges represent relationships based on both morphological similarity and spatial proximity. This hypergraph is processed by a Graph Neural Network (GNN), which incorporates hypergraph convolutions and attention mechanisms as detailed in previous work. The resulting feature vectors for each region are pooled to create a final WSI representation. This enhanced representation,

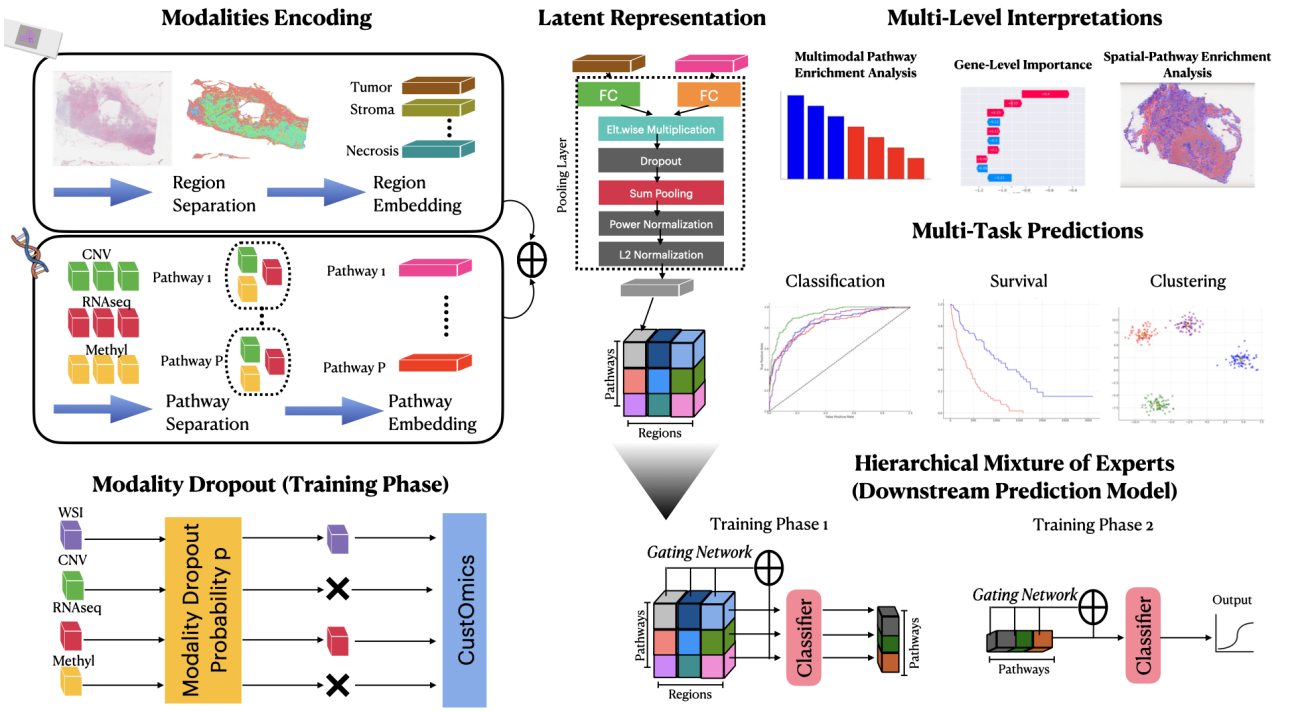


Figure 5.1: ***Modality Encoding:*** *Each modality undergoes encoding using a specific methodology. For histopathology slides, spatial regions are extracted using a hypergraph encoder to obtain an embedding for each region. Genes are segregated into gene sets in multi-omic integration, resulting in a multi-omics embedding per set.* ***Multimodal Dropout:*** *Dropout Layer for modalities to deal with missing modalities by relaxing the constraint of needing all modalities at once.* ***Latent Representation:*** *The latent representation comprises multiple blocks, each representing the embedding of the interaction between a region and a pathway.* ***Hierarchical Mixture of Experts:*** *Prediction Model based on MoE architecture on the different block embeddings. Phase 1 learns the weights of each region inside a pathway while phase 2 learns the weights of each pathway for the final prediction.* ***Multi-Task Predictions:*** *The latent representation is then utilized for supervised tasks such as classification or survival analysis or unsupervised tasks for tasks like clustering.* ***Multi-Level Interpretations:*** *Interpretation results are extracted at various levels: gene, gene-set, and spatial levels.*

based on hard-clustered region embeddings, provides a more structured and interpretable analysis of the histopathology data.

For the multi-omics integration part, we use the original CustOmics framework introduced in chapter 3. We consider S omic sources, denoted as the set $O_i = \{O_{i,s}\}_{1 \leq s \leq S}$. Each omic source is initially partitioned into P gene sets, where each set represents distinct functional properties, denoted as $O_{i,s} = \{O_{i,s,p}\}_{1 \leq p \leq P}$. For each gene set, we employ a Variational Autoencoder (VAE)-based approach for representation learning, as originally introduced in [24]. The encoding networks, $\{\mathcal{C}_p\}_{1 \leq p \leq P}$, are

designed to integrate inputs from all omic sources related to each specific gene set, producing a latent representation $z_{i,p}^{O} \in \mathbb{R}^{z_{dim}}$ through the operation $\mathcal{C}_p(O_{i,1,p}, \ldots, O_{i,S,p}) = z_{i,p}^{O}$.

The novel enhancement in our current work involves not only concatenating these representations to form the final matrix $\mathbf{O}_i^P \in \mathbb{R}^{P \times z_{dim}}$, but also incorporating a hierarchical fusion strategy. This strategy dynamically weighs the importance of each gene set across different omic sources, providing a more nuanced integration of multi-omics data. The resulting fused representation captures both the shared and unique information across the various omic sources, thereby enabling a more robust and interpretable model for downstream predictive tasks.

Multimodal Representation and Prediction

Upon acquiring the Whole Slide Image (WSI) bag $\mathbf{X}_i^K$ and the multi-omics bag $\mathbf{O}_i^P$, the subsequent phase involves constructing the final multimodal representation. This representation is crafted via a bilinear operation between each pair of elements from both bags, generating a 3-dimensional tensor $\mathbf{Z} \in \mathbb{R}^{K \times P \times z_{dim}}$ where $z_{k,p} = \mathcal{B}((\mathbf{X}_i^K)_k, (\mathbf{O}_i^P)_p)$ and $\mathcal{B}$ signifies the bilinear fusion operator.

To effectively capture the complex interactions between the multimodal features from the WSI and multi-omics data, we adopt the Multimodal Factorized Bilinear (MFB) pooling method introduced by Yu et al. [266]. MFB is designed to efficiently model second-order interactions between different feature spaces while reducing computational complexity and the number of parameters required.

The MFB method begins by projecting the feature vectors from the WSI bag $\mathbf{X}_i^K$ and the multi-omics bag $\mathbf{O}_i^P$ into a common low-dimensional space. Specifically, for a region feature vector $(\mathbf{X}_i^K)_k$ and a pathway feature vector $(\mathbf{O}_i^P)_p$, the projections are computed as follows:

$$\mathbf{x}'_k = \mathbf{W}_x(\mathbf{X}_i^K)_k \quad \text{and} \quad \mathbf{y}'_p = \mathbf{W}_y(\mathbf{O}_i^P)_p$$

where $\mathbf{W}_x \in \mathbb{R}^{k \times z_{dim}}$ and $\mathbf{W}_y \in \mathbb{R}^{k \times z_{dim}}$ are learnable weight matrices, and k is the dimension of the common space.

Next, the element-wise product of these projected vectors is computed:

$$z'_{k,p} = \mathbf{x}'_k \odot \mathbf{y}'_p$$

where $\odot$ denotes the Hadamard (element-wise) product. This operation captures the interaction between corresponding elements of the WSI and multi-omics feature vectors in the reduced-

dimensional space.

To further refine the multimodal representation, MFB employs a sum-pooling operation over m different instantiations (factors) of the linear projections. Each factor generates a vector $z'_{k,p,i}$, and the final output vector $z_{k,p}$ is obtained by summing these vectors:

$$z_{k,p} = \sum_{i=1}^{m} z'_{k,p,i}$$

This sum-pooling step aggregates the information from multiple factors, resulting in a compact and informative feature representation $z_{k,p} \in \mathbb{R}^k$. Finally, the resulting multimodal tensor $\mathbf{Z} \in \mathbb{R}^{K \times P \times z_{dim}}$ encapsulates the complex interactions between the WSI regions and the multi-omics pathways.

Once the final multimodal tensor $\mathbf{Z}$ is constructed, it is fed into a downstream network $\mathcal{D}$, responsible for producing the model's final prediction. This downstream network comprises a hierarchical mixture of expert networks, which operates in two distinct stages. In the initial stage, for each pathway p, all the regions are inputted into a mixture of experts network [129]. This network yields a pathway representation $z_p^{moe} = \sum_{k=1}^{K} w_k^p z_{p,k}$, where $(w_k^p)_k$ with $\sum_{k=1}^{K} w_k^p = 1$ are trainable parameters. These pathway representations are then individually directed into single linear layers, each responsible for producing predictions specific to their respective pathways.

In the second stage, the pathway representations are consolidated by inputting them into a second Mixture-of-Experts (MoE) network. This network aims to aggregate the pathway-specific representations into a unified representation $z^{moe} = \sum_{p=1}^{P} w_p z_p^{moe}$, where $(w_p)_p$ with $\sum_{p=1}^{P} w_p = 1$ are trainable parameters. The aggregated representation z^{moe} is then passed through a final linear layer to generate the overall model prediction.

This hierarchical approach, combined with the MFB pooling method, allows the model to effectively leverage the rich, multimodal information from both WSI and multi-omics data, ultimately enhancing the predictive power and interpretability of the model.

Multimodal Dropout

To better enforce the robustness of our method to missing data, we implement multimodal dropout introduced in Cheerla et al.[46] to deal with missing modalities under the assumption of modalities missing at random. Instead of dropping single neurons, the idea is to drop entire feature vectors

corresponding to specific modalities so that it scales up the weights of the others. This is applied to each sample data with a probability p for each modality. The dropout rate is a hyperparameter that needs to be tuned.

Downstream Task

CustOmics accommodates training for two distinct tasks. Firstly, a supervised classification task aims to predict the probability of each class occurrence. This task requires training by employing a standard categorical cross-entropy loss computed between the predicted classes and the ground truth labels.

The second task involves predicting survival outcomes, trained using the DeepSurv loss function outlined in Katzman et al. [133]. The model adopts the negative partial log-likelihood formula, expressed in our context as:

$$L(\theta) = -\sum_{i:E_i=1}\left(\hat{\mu}(x_i;\theta) - \log \sum_{j\in\mathcal{R}(T_i)} e^{\hat{\mu}(x_j;\theta)}\right) \tag{5.1}$$

where E_i represents the event for patient i, $\hat{\mu}(x;\theta)$ denotes the risk function associated with the risk score estimated by the network's output layer, and $\mathcal{R}(t)$ defines the risk set, signifying the patients still susceptible to failure after time t.

5.2.2 . Multi-level Interpretability

To make the results of the CustOmics model interpretable, we implement multiple scores to understand the predictions at different levels of the integration process.

Gene Importance & Pathway Enrichment

In pursuit of enhancing gene-level interpretability, we adapt the method introduced by Withnell et al. [256] to compute SHAP (Shapley Additive Explanations) values for deep variational autoencoders, as described in chapter 3.

After training the Multimodal CustOmics network, we compute SHAP values (similarly to chapter 3) for genes or latent dimensions within the multi-omics embedding part. SHAP values are a game-theoretic approach that attributes the contribution of each feature (in this case, each gene or latent dimension) to the prediction made by the model.

For a given patient i, let $\mathbf{z}_i$ represent the latent embedding generated by the network, where

$\mathbf{z}_i \in \mathbb{R}^{z_{dim}}$. The SHAP value $s_{i,g}$ for a gene g is calculated by averaging the marginal contributions of the gene across all possible feature subsets, defined as:

$$s_{i,g} = \sum_{S \subseteq \mathbf{z}_i \setminus \{g\}} \frac{|S|!(z_{dim} - |S| - 1)!}{z_{dim}!} \left[f(\mathbf{z}_{i,S \cup \{g\}}) - f(\mathbf{z}_{i,S}) \right]$$

where S is a subset of features excluding g, and $f(\cdot)$ represents the model's predictive function. This value measures the average contribution of the gene g to the model's prediction for patient i, when considered across all possible subsets of genes.

To generalize this to the multimodal setting, we compute SHAP values for each modality separately and then aggregate them across modalities to obtain a comprehensive view of the gene's importance. The SHAP values $s_{i,g}$ are averaged across samples with similar features, providing insights at various training phases, thereby highlighting gene importance in single-omic, multi-omic, and multimodal integration. Detailed explanations of these processes can be found in Appendix A.5.

To further enhance biological interpretability, we propose the derivation of a Pathway Enrichment Score (PES) to assess the impact of specific pathway activations on prediction tasks. This is achieved by leveraging the weights learned by the gating networks in the Mixture of Experts (MoE) model within the CustOmics framework.

For each patient i, let w_{ip} denote the weight assigned by the gating network to pathway p. The ranking score r_{ip} for patient i and pathway p is defined as:

$$r_{ip} = (w_p)_i$$

This score r_{ip} quantifies the overall contribution of pathway p to the final prediction for patient i. This pathway ranking forms the basis for subsequent pathway enrichment analysis.

At the population level, and inspired by the work of Lundberg et al. [108], we use the computed SHAP values and the pathway importance scores to conduct gene set variation analysis (GSVA). For each patient i, let $s_{i,g}$ represent the SHAP value for gene g. To generalize these values across all pathways, we normalize the SHAP values by the importance score of the pathway associated with the gene, resulting in the normalized SHAP value $\tilde{s}_{i,g}$:

$$\tilde{s}_{i,g} = r_{ip(g)} \cdot s_{i,g}$$

where $p(g)$ represents the pathway associated with gene g. Some genes can belong to multiple pathway, in this case we take a mean value of all the corresponding scores.

These normalized scores $\tilde{s}_{i,g}$ are then used as rankings for computing associated p-values for each pathway using the Kolmogorov-Smirnov (K-S) test. The K-S test compares the distribution of genes within a pathway against the distribution of all other genes, enabling the identification of pathways that are significantly enriched in the context of the patient's profile.

The resulting p-values provide a statistical measure of the impact of each pathway on the prediction task, facilitating the identification of biologically relevant pathways that contribute to the model's decisions.

Spatial-level

To evaluate the importance of the interaction between a spatial region and a functional group, we compute a Multimodal Interaction Score (MIS). This score is directly obtained from the weights of the gating network such that for a specific patient i, we have $MIS_{i,k,p} = (w_k^p)_i$. The score measures the impact of a multimodal interaction between a spatial region and a functional group on the final prediction.

5.2.3 . Experimental Setup

Dataset Description

This study uses the pan-cancer dataset from the Genomic Data Commons (GDC) [100], comprising 11,768 patients across 33 tumor types and encompassing multi-omics data, histopathology slides, and clinical data. We assess the performance of our model on both the entire pan-cancer dataset and smaller cohorts of specific tumor types to demonstrate the robustness of our approach concerning varying patient numbers. The objective is to evaluate CustOmics for tumor type classification and survival outcome prediction. For the survival prediction task, we selected eight cohorts based on patient numbers and censoring rates, as outlined in Table S1. Three of these eight cohorts were also utilized to assess classification into molecular subtypes: TCGA-BRCA, TCGA-COAD, and TCGA-STAD.

Implementation Details

The CustOmics framework is based on the Pytorch deep-learning library [192]. It can be applied to any combination of high-dimensional datasets and histopathology images with multitask training. As done in *Zhang et al.* [273], DNA methylation data can be divided into 23 separate blocks, each feeding

a hidden layer corresponding to a chromosome to avoid overfitting and save GPU memory.

Inspired by the work in [24], we adopt a multiphase training strategy to ensure the optimal integration of all modalities. During the first phase, we train each modality independently to obtain unsupervised sub-representations for the different bags. The second phase consists of unfreezing the central encoding network that learns the supervised representation of the final bag.

The whole architecture is built using fully connected blocks with weights initialized following a uniform distribution $\mathcal{U}(-\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{k}})$ where k is the number of weight parameters. We use a batch normalization technique in each layer composing the neural network to address the internal covariate shift problem[121]. Also, to avoid overfitting problems, we use dropout [228]; its rate is considered a hyperparameter.

The input dataset was randomly split into training, validation, and testing sets (60-20-20%) using stratified 5-fold cross-validation so that the proportion of samples in each tumor type between the different sets is preserved in all the folds. We perform Bayesian optimization [227] using the validation set to find our model's best possible combination of hyperparameters.

5.3 . Results

5.3.1 . Prediction Results

We executed multiple test cases in a comprehensive evaluation of CustOmics within a multitask framework encompassing classification and survival analyses. The performances across survival and classification tasks are presented in Tables 5.1 and 5.2. We initially assessed cancer-type classification within the TCGA Pancancer cohort, revealing CustOmics' superior performance in terms of AUC compared to other benchmarked methods [107]. This notable performance can be attributed to the substantial patient cohort size and the rich information embedded within molecular data, a phenomenon well-documented across multiple studies [274, 273, 24].

To underscore the method's robustness concerning sample size, we evaluated smaller datasets, focusing on predicting molecular subtypes within three specific TCGA cohorts: TCGA-BRCA, TCGA-COAD, and TCGA-STAD. Across all classification tasks (as detailed in Table 5.1), CustOmics consistently outperformed other comparable methods. Notably, in multi-omics integration, the mixed-integration VAE within CustOmics demonstrated superior performance compared to the SNN utilized by other

methods. An additional assessment showcased the significance of multi-omics integration, delineated in Table 5.3, emphasizing the impact on performance when replacing the VAE encoder in CustOmics with a standard SNN. Further exploration into the integration strategies revealed the advantage of CustOmics in exploiting diverse molecular data sources. Comparative analyses in Table 5.4 demonstrated that while SNN performed best with RNAseq alone, CustOmics exhibited enhanced performance when integrating RNAseq with CNV and methylation data. This divergence in integration strategies suggests CustOmics' capacity to augment predictive power and unveil novel interactions among disparate data sources.

The evaluation extended to survival outcome prediction across eight TCGA cohorts, consistently showcasing CustOmics' superior predictive capabilities compared to alternative methods.

Notably, in survival analysis based solely on Whole Slide Images (WSI), CustOmics exhibited comparatively weaker performance in specific cohorts than the transformer architecture employed in MCAT. An ablation study (detailed in Table S4) replaced CustOmics' hypergraph encoding for WSI embeddings with a visual transformer. While this configuration works better for WSI only, CustOmics yields superior performance with the hypergraph embeddings for multimodal representation learning. Despite the inferior results in a WSI-only setting, those results show that a hypergraph-based embedding is better suited for multimodal integration than visual transformers.

In survival tasks, particularly in multi-omics scenarios (without WSI), CustOmics displayed substantially more significant differences in performances, especially as other methods like SNN showed concordance indices approaching randomness (Table 5.2). CustOmics' lower standard deviation across folds underscores its enhanced robustness compared to other state-of-the-art approaches.

5.3.2 . Multi-level Explainability: Classification

CustOmics can conduct pathway enrichment analysis across multiple tasks. Figure 5.2 presents interpretability findings concerning the PAM50 subtype classification within the TCGA-BRCA dataset. The objective is to elucidate the determinants driving the discrimination of specific subtypes, notably the Her2 subtype, within a multimodal context.

The initial layer of interpretability operates at the gene level, utilizing a Multi-Omics Pathway Enrichment Score (MPES) and conducting Gene Set Variation analysis using normalized gene importance scores (details in the methods section). Figure 5.2b delineates essential pathways in Her2 subtype discrimination, notably highlighting the significance of estrogen response and KRAS signaling down

Table 5.1: ***Classification Performances:*** *Comparison of the classification performances for the 4 tasks with respect to the area under ROC (AUC %).*

Methods	**PANCAN**	**BRCA**	**COAD**	**STAD**
SNN (Multi-Omics)	94.1 ± 2.7	92.0 ± 3.3	79.2 ± 3.3	84.6 ± 3.3
CustOmics (Multi-Omics)	$\mathbf{98.9 \pm 1.4}$	$\mathbf{98.3 \pm 1.0}$	$\mathbf{88.1 \pm 1.9}$	$\mathbf{98.4 \pm 1.2}$
DeepSets (WSI Only)	84.7 ± 3.3	68.6 ± 4.1	55.2 ± 4.3	58.9 ± 4.6
AttnMIL (WSI Only)	88.4 ± 2.0	71.2 ± 4.9	56.2 ± 4.4	61.4 ± 4.4
DeepAttnMISL (WSI Only)	89.8 ± 2.5	71.1 ± 3.3	55.7 ± 4.0	62.1 ± 4.5
MCAT (WSI Only)	90.4 ± 1.8	72.3 ± 3.3	61.7 ± 3.3	69.4 ± 3.5
CustOmics (WSI Only)	$\mathbf{92.5 \pm 1.2}$	$\mathbf{73.2 \pm 3.1}$	$\mathbf{62.2 \pm 2.0}$	$\mathbf{71.4 \pm 2.2}$
DeepSets (WSI + Multi-Omics)	96.7 ± 1.5	84.9 ± 2.0	58.6 ± 2.2	67.1 ± 2.7
AttnMIL (WSI + Multi-Omics)	97.1 ± 1.2	86.6 ± 2.1	60.0 ± 2.5	69.4 ± 2.7
DeepAttnMISL (WSI + Multi-Omics)	97.8 ± 1.1	88.3 ± 2.7	65.4 ± 2.7	66.6 ± 2.2
MCAT (WSI + Multi-Omics)	98.9 ± 1.1	95.4 ± 2.0	93.3 ± 1.2	88.7 ± 2.3
CustOmics (WSI + Multi-Omics)	$\mathbf{99.5 \pm 0.9}$	$\mathbf{98.7 \pm 1.1}$	$\mathbf{94.7 \pm 1.0}$	$\mathbf{96.3 \pm 2.4}$

Table 5.2: ***Survival Performances:*** *Comparison of the survival performances for the 8 TCGA cohorts with respect to the Concordance Index (C-index %).*

Methods	**BLCA**	**BRCA**	**COAD**	**GBMLGG**	**KIRC**	**LUAD**	**STAD**	**UCEC**
SNN (Multi-Omics)	54.6 ± 2.7	47.5 ± 4.9	50.8 ± 3.4	60.5 ± 3.1	59.0 ± 5.7	54.2 ± 5.4	51.1 ± 4.6	49.6 ± 8.3
CustOmics (Multi-Omics)	$\mathbf{64.1 \pm 1.5}$	$\mathbf{63.4 \pm 1.8}$	$\mathbf{58.6 \pm 1.8}$	$\mathbf{78.4 \pm 1.6}$	$\mathbf{66.5 \pm 2.3}$	$\mathbf{63.7 \pm 1.2}$	$\mathbf{55.6 \pm 1.4}$	$\mathbf{68.5 \pm 2.8}$
DeepSets (WSI Only)	50.6 ± 5.1	50.1 ± 6.5	50.2 ± 3.9	50.9 ± 3.5	49.6 ± 5.4	50.7 ± 7.5	50.0 ± 7.5	50.3 ± 7.5
AttnMIL (WSI Only)	54.8 ± 4.8	57.0 ± 5.7	59.6 ± 3.9	79.0 ± 3.2	56.9 ± 6.1	56.3 ± 6.8	57.8 ± 6.1	$63. \pm 7.$
DeepAttnMISL (WSI Only)	50.3 ± 5.5	52.1 ± 5.9	53.1 ± 3.3	73.8 ± 3.9	56.9 ± 6.6	55.0 ± 6.1	58.8 ± 5.3	60.0 ± 7.2
MCAT (WSI Only)	55.5 ± 3.2	57.1 ± 5.6	$\mathbf{59.9 \pm 2.5}$	$\mathbf{79.4 \pm 2.0}$	56.0 ± 3.4	55.0 ± 6.1	57.7 ± 3.9	63.4 ± 6.7
CustOmics (WSI Only)	$\mathbf{56.7 \pm 3.9}$	$\mathbf{59.4 \pm 3.7}$	58.5 ± 2.1	78.7 ± 2.0	$\mathbf{61.3 \pm 2.2}$	$\mathbf{60.0 \pm 5.5}$	$\mathbf{59.6 \pm 2.3}$	$\mathbf{67.9 \pm 2.2}$
DeepSets (WSI + Multi-Omics)	59.6 ± 4.7	52.1 ± 7.1	61.4 ± 3.6	81.8 ± 3.3	54.2 ± 5.4	56.8 ± 7.3	52.9 ± 5.8	59.0 ± 7.4
AttnMIL (WSI + Multi-Omics)	57.4 ± 5.3	54.8 ± 6.4	61.9 ± 3.5	81.3 ± 3.0	62.0 ± 6.2	58.5 ± 6.7	54.0 ± 6.2	56.9 ± 6.1
DeepAttnMISL (WSI + Multi-Omics)	58.5 ± 5.4	58.0 ± 7.6	61.0 ± 3.2	81.4 ± 3.3	60.7 ± 7.1	55.0 ± 6.1	53.8 ± 5.7	59.1 ± 6.6
MCAT (WSI + Multi-Omics)	62.4 ± 3.6	58.3 ± 5.5	62.8 ± 2.9	82.5 ± 2.3	66.5 ± 3.9	62.5 ± 4.5	56.2 ± 3.1	62.2 ± 2.7
CustOmics (WSI + Multi-Omics)	$\mathbf{67.2 \pm 2.5}$	$\mathbf{65.2 \pm 3.6}$	$\mathbf{64.5 \pm 2.2}$	$\mathbf{84.2 \pm 2.3}$	$\mathbf{68.2 \pm 2.1}$	$\mathbf{64.9 \pm 3.7}$	$\mathbf{58.0 \pm 1.5}$	$\mathbf{68.0 \pm 2.2}$

hallmarks. The interrelation between the Her2 subtype and estrogen response has been extensively investigated [172], emphasizing their coexpression's multifaceted impact on breast carcinogenesis, invasive behavior, and cellular growth.

Further exploration into gene-level importance is depicted in Figure 5.2c, spotlighting the predominant genes responsible for discriminating the Her2 subtype and contrasting their importance across other subtypes. Notably, the FOXA1 gene emerges with substantial importance, aligning with its suggested role as a transcription factor for Her2, as indicated in Cruz et al. [61]. Beyond multi-omics pathway enrichment analysis, CustOmics extends interpretability to encompass multimodal enrichment, revealing spatial interactions within histopathology slides that correlate with specific pathways

Table 5.3: ***Ablation Study*** *Performance comparison between CustOmics and the state of the art for classification tasks by replacing different instances of the model:* ***a. Multi-Omics Ablation:*** *CustOmics A1 replaces the multi-omics VAE with an SNN.* ***b. Hypergraph Ablation:*** *CustOmics A2 replaces the hypergraph encoder with a visual transformer and CustOmics A3 replaces the hypergraph representation with a regular graph embedding.* ***c. Downstream Network Ablation:*** *CustOmics A4 replaces the hierarchical mixture-of-experts approach with a regular mixture-of-experts network, CustOmics A5 replaces it with a transformer classifier.*

Methods	PANCAN	BRCA	COAD	STAD
SNN (Multi-Omics)	94.1 ± 2.7	92.0 ± 3.3	79.2 ± 3.3	84.6 ± 3.3
CustOmics (Multi-Omics)	$\mathbf{98.9 \pm 1.4}$	$\mathbf{98.3 \pm 1.0}$	$\mathbf{88.1 \pm 1.9}$	$\mathbf{98.4 \pm 1.2}$
DeepSets (WSI Only)	84.7 ± 3.3	68.6 ± 4.1	55.2 ± 4.3	58.9 ± 4.6
AttnMIL (WSI Only)	88.4 ± 2.0	71.2 ± 4.9	56.2 ± 4.4	61.4 ± 4.4
DeepAttnMISL (WSI Only)	89.8 ± 2.5	71.1 ± 3.3	55.7 ± 4.0	62.1 ± 4.5
MCAT (WSI Only)	90.4 ± 1.8	72.3 ± 3.3	61.7 ± 3.3	69.4 ± 3.5
CustOmics A1 (WSI Only)	85.1 ± 3.8	70.5 ± 5.5	57.7 ± 4.1	59.4 ± 4.4
CustOmics A2 (WSI Only)	90.4 ± 1.8	72.3 ± 3.3	61.7 ± 3.3	69.4 ± 3.5
CustOmics A3 (WSI Only)	88.7 ± 1.2	69.4 ± 3.6	55.1 ± 3.5	61.9 ± 5.0
CustOmics A4 (WSI Only)	89.4 ± 2.1	71.0 ± 3.3	58.8 ± 3.3	67.2 ± 3.7
CustOmics A5 (WSI Only)	87.1 ± 1.9	70.4 ± 3.6	55.2 ± 3.9	66.4 ± 4.1
DeepSets (WSI + Multi-Omics)	96.7 ± 1.5	84.9 ± 2.0	58.6 ± 2.2	67.1 ± 2.7
AttnMIL (WSI + Multi-Omics)	97.1 ± 1.2	86.6 ± 2.1	60.0 ± 2.5	69.4 ± 2.7
DeepAttnMISL (WSI + Multi-Omics)	97.8 ± 1.1	88.3 ± 2.7	65.4 ± 2.7	66.6 ± 2.2
MCAT (WSI + Multi-Omics)	98.9 ± 1.1	95.4 ± 2.0	93.3 ± 1.2	88.7 ± 2.3
CustOmics A1 (WSI + Multi-Omics)	96.1 ± 1.7	85.2 ± 2.4	59.5 ± 1.9	67.7 ± 2.9
CustOmics A2 (WSI + Multi-Omics)	99.0 ± 1.3	98.4 ± 2.2	94.1 ± 1.2	95.3 ± 2.4
CustOmics A3 (WSI + Multi-Omics)	99.9 ± 1.1	98.2 ± 2.5	93.9 ± 1.9	94.1 ± 2.7
CustOmics A4 (WSI + Multi-Omics)	99.4 ± 0.8	98.0 ± 1.2	94.2 ± 1.1	96.1 ± 2.2
CustOmics A5 (WSI + Multi-Omics)	98.5 ± 1.9	97.3 ± 3.4	93.8 ± 2.0	94.2 ± 2.9

Table 5.4: ***Modality Combinations*** *Performance comparison between multiple combination of modalities for CustOmics for the Pancancer classification task. The evaluation is done using the Area Under ROC-curve (AUC).*

Omics Combinations	SNN	CustOmics
CNV	74.3 ± 3.0	75.1 ± 2.7
RNAseq	94.0 ± 2.6	96.0 ± 1.4
Methyl	81.1 ± 1.7	82.3 ± 1.3
CNV + RNAseq	94.3 ± 2.9	96.9 ± 0.8
CNV + Methyl	81.4 ± 1.8	85.7 ± 2.1
RNAseq + methyl	93.2 ± 1.0	97.3 ± 0.7
CNV + RNAseq + Methyl	94.1 ± 2.7	98.9 ± 1.4

for discriminating Her2 subtypes. Figure 5.2d showcases such interpretability outcomes for the estrogen response and KRAS pathways. Different cell populations within high-importance regions for

each pathway are described to help biological interpretation. Notably, regions associated with the KRAS pathway exhibit increased proportions of stromal cells, suggesting potential regulation of tumor cell signaling via these stromal cells [235]. Conversely, the estrogen response demonstrates a strong interaction with regions featuring elevated densities of Tumor-Infiltrating Lymphocytes (TILs) and lymphocytes, corroborated by multiple sources [71, 156]. This interaction holds particular significance when stratifying between ER- and ER+ patients.

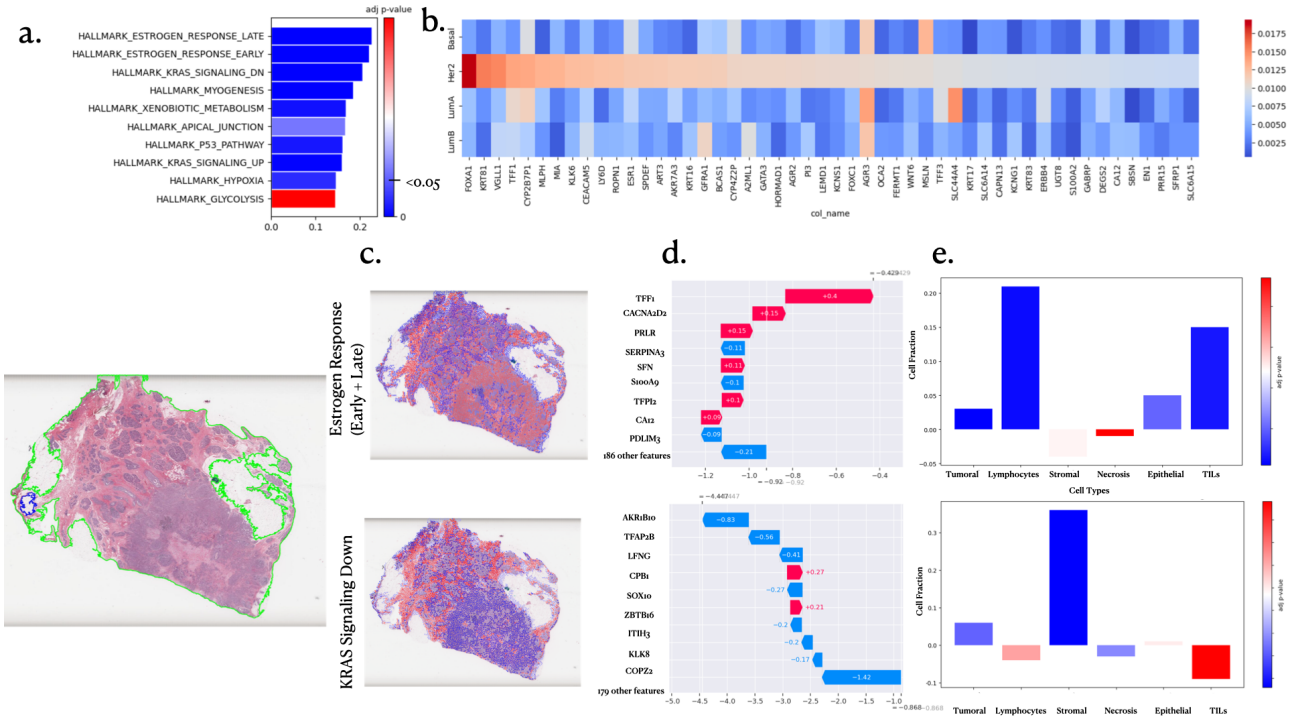


Figure 5.2: ***PAM50 Explainability Analysis: a.*** *Pathway enrichment scores and associated p-values from the gene set variation analysis.* ***b.*** *SHAP values for the most influential genes affecting the stratification of the Her2 subtype and their impact on other subtypes.* ***c.*** *Spatial Enrichment Analysis for the top 2 pathways and their key genes.* ***d.*** *Gene importance within the considered pathways.* ***e.*** *Cell distribution in the top 10% attention regions.*

5.3.3 . Multi-level Explainability: Survival

In a similar vein, interpretability analysis extends to survival analysis. Figure 5.3 delineates the varying degrees of enrichment analysis for predicting survival outcomes within the TCGA Pancancer dataset.

Specifically, Figure 5.3b underscores the predominant influence of the inflammatory response

pathway on survival outputs, a finding consistent with existing literature [275].

Demonstrating the relevance of employing multimodal integration in pan-cancer survival analysis, Figure 5.3c illustrates the impact of incorporating multiple data sources on stratifying low and high-risk patients, as evidenced by Kaplan-Meier curves and their corresponding log-rank p-values.

Further investigation into the effect of essential pathways is portrayed in Figure 5.3d, showcasing the significance of interactions between the two most crucial pathways. Notably, heightened importance is observed within the inflammatory response pathway, characterized by increased lymphocyte densities and Tumor-Infiltrating Lymphocytes (TILs). In contrast, the epithelial-mesenchymal transition pathway manifests greater densities of stromal cells.

Delving deeper into the Epithelial-Mesenchymal Transition pathway, the primary influential gene appears to be FBN2, renowned for its inhibition of cancer cell invasion and migration, as stated in Mahdizadehi et al. [163], thereby explaining its inclination toward lower risk outcomes.

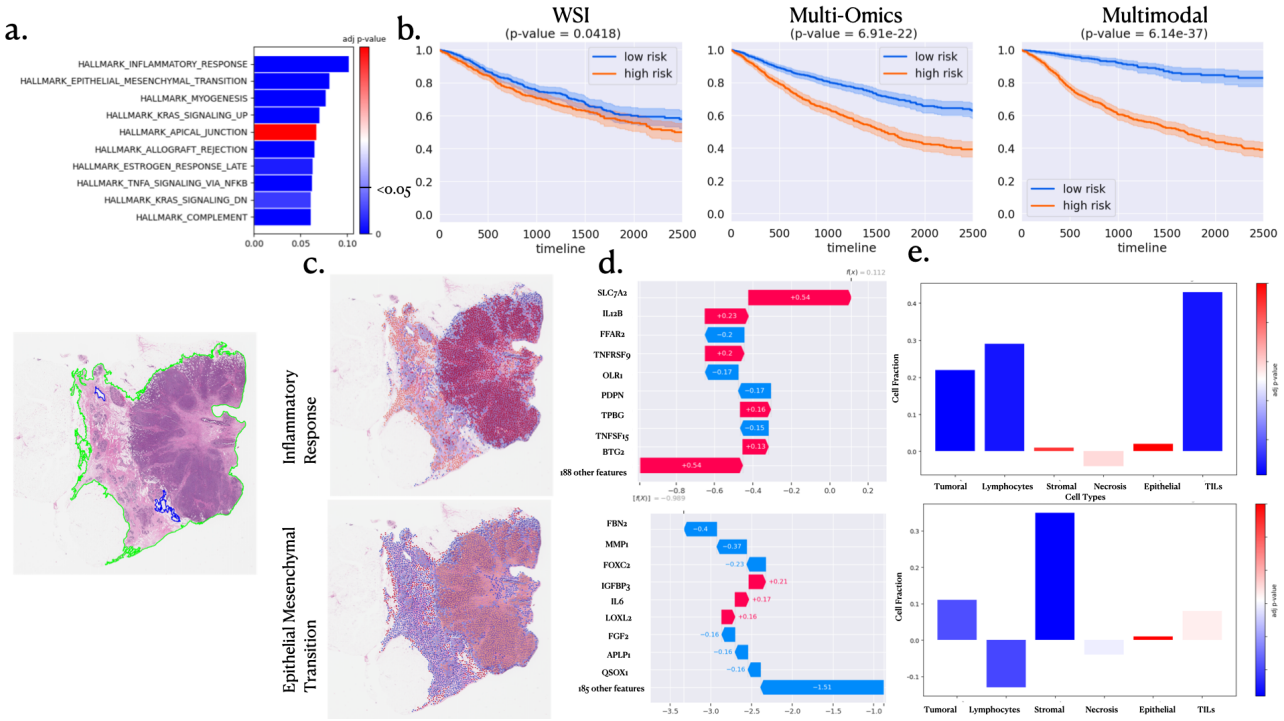


Figure 5.3: ***Explainability Analysis for Pan-Cancer Survival Outcome Prediction Task:*** ***a.*** *Pathway enrichment analysis.* ***b.*** *Kaplan Meier curve associated with the survival outcome prediction task, showing the high and low risk for death event stratification with a computed log-rank p-value.* ***c.*** *Spatial Enrichment Analysis for the top 2 pathways and their most important genes.* ***d.*** *Gene importance within the considered pathways.* ***e.*** *Cell distribution in the top 10% attention regions.*

5.3.4 . Application to the Integration of Multi-Omics Data & Histopathology Slides for Survival Analysis in Lung Cancer

Lung cancer remains one of the most prevalent and deadly forms of cancer worldwide, posing significant challenges to effective diagnosis and treatment. Its complex molecular landscape and varied histopathological features necessitate advanced strategies for accurate characterization and personalized therapeutic approaches. Traditional methods often fall short of capturing the complexity of lung cancer, highlighting the need for more comprehensive and integrative methodologies. This short study serves as a validation for our method's explainability results.

This study uses the International Adjuvant Lung Cancer Trial (IALT) dataset. It is a comprehensive collection of clinical data from large-scale, randomized controlled trials to evaluate adjuvant chemotherapy's efficiency in patients with resected non-small cell lung cancer (NSCLC). We study a subset of this trial comprised of 544 patients with histopathology slides, mutation status, and Copy Number Variation (CNV) data. A detailed description of the dataset is presented in Appendix B.3.

We performed survival outcome prediction on the TCGA and IALT cohorts to validate explainability results. The comprehensive analysis of C-index values for different models, as shown in Figure 5.4b, further illustrates the superiority of multimodal approaches integrating omics and whole slide images (WSI). Specifically, the multimodal model significantly outperformed individual omics or WSI models in both datasets, suggesting that combining these data types can enhance predictive accuracy for survival outcomes.

The pathway enrichment analysis, illustrated in Figure 5.4, revealed congruent outcomes between the TCGA and IALT datasets, identifying KRAS signaling down as the foremost pathway. This finding aligns with established literature [232], which underscores the prevalence of oncogenic KRAS mutations in approximately 25%

Subsequently, we explored the interactions between the KRAS signaling-down pathway and spatial regions within histopathology images. Figure 5.4b depicts that regions of high attention in both datasets exhibit similar distributions of cell types, indicating the robustness of our method across distinct datasets of the same cancer type. Notably, this distribution underscores the association between the KRAS pathway and tumor-infiltrating lymphocytes (TILs) heightened densities and marginally increased tumoral cell counts in predicting survival outcomes. This consistent association echoes previous findings in [154], highlighting a solid correlation between KRAS mutation status and tumor

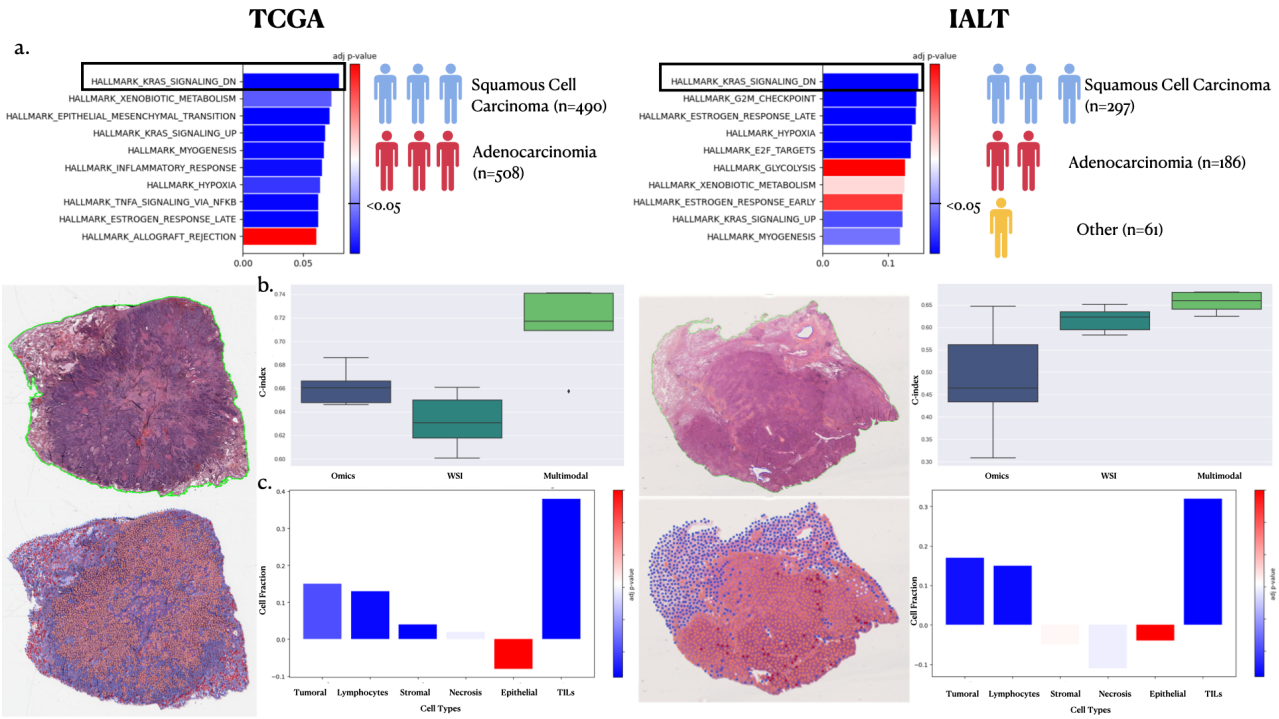


Figure 5.4: *Comparison of interpretations between the TCGA-LUAD/TCGA-LUSC datasets and IALT regarding the distinction between high and low survival risk.* ***a.*** *Pathway Enrichment analysis highlights the task's top pathways.* ***b.*** *Spatial Importance of the KRAS signaling-down pathway, illustrating regions of high importance in interaction with this pathway.* ***c.*** *Comparison of cell populations within the Whole Slide Image (WSI) high-importance regions.*

immunity-related characteristics, notably CD8+ TILs.

5.4 . Discussion

The CustOmics framework is a comprehensive toolset to bridge prediction and interpretation within biological systems across multiple levels: genes, pathways, and spatial orientations. This multifaceted system generates three distinct interpretability scores concurrent with predictions, unraveling the biological knowledge underlying model outcomes. Empirical assessments underscore CustOmics' robust predictive capabilities, outperforming state-of-the-art methodologies in integrating multi-omics and histopathology data across eight diverse datasets. However, despite its efficiency with smaller datasets in this study, CustOmics' reliance on deep learning methodologies might restrict efficiency when confronted with limited training data availability.

Furthermore, CustOmics stands out for its interpretability, facilitating a broader spectrum of analyses and enriching understanding across diverse biological modalities. Notably, although this study centered on three omics data types, CustOmics exhibits versatility in seamlessly integrating varied omics data without necessitating framework alterations. This adaptability originates from an initial phase that independently trains on each source, serving as a normalization layer for heterogeneous sources.

The strategic segmentation of inputs into interpretable entities for spatial and molecular data mitigates challenges arising from the high dimensionality of whole slide images and multi-omics datasets. This partitioning augments interpretability and broadens the method's applicability to uncharted pathways or spatial regions beyond this study's scope.

CustOmics places significant emphasis on expansive interpretability functionalities. This aims to unveil predominant biological functions steering specific predictions across diverse data sources and scales. This comprehensive approach fosters collaboration between biologists and computational pathologists, offering a framework for in-depth analyses through enrichment analysis for omics and spatial data. The method extracts coherent insights from diverse data sources, unveiling a panoramic view of interconnected biological processes influencing the outcome of interest. By integrating omics and spatial data within enrichment analysis, CustOmics enables a deeper understanding of the interplay between molecular information and spatial contexts, enriching investigative pathways for researchers in the field.

Despite its potential and performance, our method has a few noteworthy limitations. Firstly, while this study successfully delineates interactions between omics and histopathology data, a method to effectively discern and capitalize on the individual contributions of each omic source to a patient's molecular profile still needs to be present. It prevents a comprehensive understanding of each omic source's distinct impact on the molecular landscape. Secondly, the link established between the generated representation and phenotype data, beyond mere predictive labels, solely relies on the conditioning of the latent space. While this conditioning methodology effectively incorporates phenotypical signals into the multimodal representation, it lacks a mechanism to unveil interactions between different modalities and diverse clinical variables explicitly. This omission presents an avenue for future development, potentially enhancing the interpretability of multimodal interactions and their associations with clinical factors.

Conclusions & Perspectives

Contents

Abstract

The conclusion of this thesis synthesizes the advancements made in integrating multi-omics data and histopathology for precision oncology. It highlights the challenges of multimodal data integration, particularly the high-dimensional nature of bulk omics and whole-slide imaging (WSI). The role of spatial transcriptomics is briefly discussed as a powerful tool for enhancing the spatial resolution of molecular data, particularly when combined with WSI and multi-omics bulk data. The concluding section also addresses future perspectives, including the potential of deep learning and foundation models in overcoming the challenges of multimodal data integration and advancing personalized cancer treatment.

6.1 . High-Dimensional Multimodal Representation Learning

The work presented in this thesis introduces innovative methodologies to address the challenges associated with the integration and analysis of high-dimensional multimodal data, particularly in the domain of precision oncology. The integration of multi-omics data and histopathology enables researchers to gain a comprehensive understanding of the complex biological processes that drive cancer, yet poses significant technical and analytical challenges due to the scale, complexity, and differing nature of the data involved. The primary contributions of this thesis center on overcoming these challenges through the development of novel computational frameworks: CustOmics and a hypergraph representation for whole-slide images (WSIs) and a mixture of both through Multimodal CustOmics.

6.1.1 . Interest in Multimodal Data Integration for Oncology

The integration of multiple data modalities—such as genomic, transcriptomic, proteomic, and histopathological data—holds transformative potential in cancer research. Each of these modalities provides unique insights into the biological processes at play, but when studied independently, important interactions and relationships between different levels of biological information may be missed. For instance, omics data provide detailed molecular signatures of the tumor, while histopathology provides spatial context and visual insights into the cellular and tissue architecture.

The work in this thesis is of particular interest because it goes beyond simply combining these data types. It seeks to integrate them in a way that preserves biological relevance and captures complex,

non-linear relationships between the modalities, leading to more meaningful biological interpretations. The contributions of this research are crucial for making precision oncology more actionable by providing the tools needed to analyze complex datasets and derive insights that are more reflective of the intricate nature of cancer biology.

6.1.2 . CustOmics: Capturing Complex Interactions in Multi-Omics Data

One of the core contributions of this thesis is the development of **CustOmics**, a tailored framework designed specifically for the high-dimensional nature of multi-omics data. Unlike traditional methods, CustOmics is built to handle the complexity and variability of multi-omics datasets, which often include thousands of features that capture different layers of biological information.

The key strength of CustOmics lies in its ability to preserve biological relevance. CustOmics employs advanced machine learning techniques, such as non-linear models and deep learning architectures, to ensure that the learned representations capture biologically meaningful relationships across different omics layers. This is critical in avoiding the simplifications that often arise from traditional linear integration methods.

In addition, CustOmics is designed to overcome the challenges associated with high-dimensional data. By using techniques such as dimensionality reduction with Variational Autoencoders, CustOmics reduces the complexity of the data while retaining key biological signals. This allows for the extraction of actionable insights from large datasets without succumbing to overfitting or losing critical information.

Furthermore, CustOmics enhances integration across multiple omics layers by learning shared latent spaces that reflect the common biological processes between the different types of omics data. This multi-layered integration provides researchers with the ability to explore deeper interactions between molecular features, resulting in more comprehensive insights into cancer biology.

The biological explainability of the CustOmics framework ensures that the results can be directly applied to understand cancer progression, heterogeneity, and treatment responses. By addressing the limitations of traditional integration techniques, CustOmics represents a significant advancement in how high-dimensional multi-omics data is utilized for personalized medicine.

6.1.3 . Hypergraph Representation for Whole-Slide Images (WSIs): Preserving Spatial and

Community Information

In addition to multi-omics data, histopathology data, particularly whole-slide images (WSIs), play a pivotal role in understanding tumor structure and cellular morphology. However, analyzing these large, complex images presents significant challenges, as traditional methods often fail to fully capture the spatial relationships and contextual information that are critical for understanding tumor biology.

To address this issue, a key contribution of this thesis is the development of a hypergraph-based representation for WSIs. This method is innovative in several ways. First, unlike traditional graph representations that rely on pairwise relationships between image features, the hypergraph approach captures higher-order interactions. These interactions are crucial for representing community structures and spatial relationships between different regions of the tissue, which are often essential for identifying pathological features such as immune cell infiltration or stromal architecture.

Furthermore, the hypergraph representation preserves the spatial integrity of the tissue by encoding the geometric and spatial relationships between different regions of the WSI. This is particularly important in cancer research, where the spatial arrangement of cells and tissues provides crucial insights into tumor progression and microenvironmental interactions.

Additionally, the hypergraph model is computationally efficient and scalable, making it suitable for large-scale histopathological datasets. Its robustness in maintaining both structural and spatial integrity ensures that the analyses remain biologically meaningful and applicable to a variety of clinical and research contexts.

The introduction of the hypergraph representation method is a major contribution of this thesis, as it allows for a more detailed analysis of WSIs in a way that enhances integration with molecular data. This provides a more complete picture of the tumor, linking molecular alterations with the physical structures observed in the tissue.

6.1.4 . Synthesis: Integrating Multi-Omics and Histopathology Data

The synthesis of CustOmics and the hypergraph representation for WSIs creates a powerful framework for multimodal data integration in precision oncology. The combination of these methodologies addresses the critical need for tools capable of handling both the high-dimensional nature of multi-omics data and the spatial complexity of histopathology. This integration offers several important benefits.

Firstly, the integrated framework provides cross-modal biological insights by allowing researchers to connect molecular alterations detected in multi-omics data with specific spatial features observed in histopathology. For example, CustOmics can identify genetic mutations or transcriptomic signatures associated with aggressive cancer phenotypes, while the hypergraph model can localize these alterations to specific regions of the tumor, such as the invasive front or areas of immune infiltration.

Secondly, this framework enhances the understanding of tumor heterogeneity by combining molecular and spatial data. The ability to link molecular changes with specific regions of tissue facilitates the identification of distinct subpopulations within the tumor, which may respond differently to therapy. This is a crucial step toward the development of more targeted and personalized treatment strategies.

Finally, the integration of multi-omics data and WSIs has direct clinical relevance. It can improve diagnostic accuracy, guide treatment decisions, and enhance prognostic assessments. The framework developed in this thesis provides a foundation for translating complex multimodal data into actionable insights for clinical practice, ultimately contributing to more effective and personalized cancer care.

6.1.5 . Contributions to the Field of Precision Oncology

The contributions of this work are significant both from a methodological and an applied perspective. Methodologically, the introduction of CustOmics and the hypergraph representation represents a major step forward in the field of multimodal data integration. These methods address the key challenges of high-dimensionality, biological relevance, and computational scalability, and provide solutions that are both innovative and effective.

From an applied perspective, the ability to integrate multi-omics data with histopathology helps offer a more holistic understanding of cancer biology. This work not only enhances the ability to study cancer heterogeneity and progression but also has the potential to impact clinical practice by providing more accurate diagnostics and better-informed treatment strategies.

6.2 . Explainability

In the field of precision oncology, ensuring that complex computational models and analyses are interpretable is of critical importance. Clinicians and researchers rely not only on the accuracy of predictions but also on their understanding of how and why a model arrived at a particular conclusion.

This thesis contributes significantly to the field by developing methods that prioritize explainability at various levels of analysis. The methods presented in this work ensure that the outputs of high-dimensional multimodal data integration can be understood by human experts, from molecular-level features to tissue-level spatial patterns. Key contributions include the development of H&Explainer, GMM-CeFlow, as well as the explainability features of the CustOmics framework and the hypergraph representation for WSIs.

6.2.1 . The Critical Role of Explainability in Precision Medicine

Explainability is a fundamental requirement in precision medicine, especially in oncology, where decisions regarding diagnosis, treatment, and prognosis must be made with a high degree of confidence. The integration of multimodal data—such as genomics, transcriptomics, proteomics, and histopathology—often involves complex computational models that act as "black boxes." While these models can achieve high predictive accuracy, their lack of transparency can limit their utility in clinical practice.

This thesis addresses this challenge by developing methods that explicitly focus on making complex data analysis and model outputs interpretable. These methods allow medical professionals to not only understand the predictions but also trust the underlying logic, enhancing their ability to make informed, data-driven decisions.

6.2.2 . H&Explainer: Human-Interpretable Analysis of Whole-Slide Images

A significant contribution of this thesis is the development of **H&Explainer**, a tool specifically designed to enhance the interpretability of whole-slide images (WSIs) used in histopathology. Traditional deep learning models used for WSI analysis often lack interpretability, making it difficult for pathologists to understand how predictions are derived from complex visual data.

H&Explainer addresses this issue in several ways. First, it decomposes the whole-slide images into clinically relevant components. Instead of treating WSIs as a monolithic image, H&Explainer breaks them down into meaningful regions that are directly relevant to clinicians, such as areas of immune infiltration or necrosis. Additionally, it provides layered explainability by offering interpretability at both the micro (cell-level) and macro (tissue-level) perspectives, allowing clinicians to explore model outputs in a way that aligns with their own expertise. Finally, H&Explainer translates model outputs into human-readable formats by generating visual and textual explanations that correspond to clin-

ical interpretation, making the predictions of complex models accessible and actionable for medical professionals.

6.2.3 . GMM-CeFlow: Counterfactual Analysis for Explainable Histopathology

This thesis also introduces GMM-CeFlow, a novel method designed to enhance explainability through counterfactual analysis. Counterfactual explanations are vital for understanding how small changes in features could lead to different outcomes, which is particularly important in precision oncology.

Key features of GMM-CeFlow highlight its strengths in this regard. First, it generates human-interpretable features by using Gaussian Mixture Models (GMMs) to produce features that pathologists can easily interpret and relate to clinical outcomes. In addition, GMM-CeFlow produces clinically relevant counterfactuals by generating scenarios directly linked to real-world clinical questions, such as how altering cell density or architecture could impact treatment predictions. Lastly, GMM-CeFlow supports actionable decision-making by offering insights into how specific morphological changes influence outcomes, helping clinicians explore various treatment pathways and make more informed decisions based on the analysis.

6.2.4 . Explainability in CustOmics: Understanding Multi-Omics Integration

While CustOmics is primarily designed to integrate and analyze high-dimensional multi-omics data, it places a strong emphasis on explainability to ensure that the insights it generates are accessible and useful in both clinical and research contexts.

CustOmics offers several key explainability features. One of the most important is its feature-level interpretability. CustOmics uses dimensionality reduction techniques that preserve biologically meaningful features, ensuring that critical omics elements, such as key genetic mutations or disease-driving pathways, are clearly highlighted and presented in a way that can be easily interpreted by researchers and clinicians. Another essential aspect is its multi-modal contribution analysis, which provides insights into how each omics layer, such as genomics or transcriptomics, influences the overall model predictions. This allows for a transparent understanding of how various molecular layers interact and contribute to the final outcomes, offering a clear and interpretable view of the data integration process.

6.2.5 . Explainability in the Hypergraph Representation for WSIs

The hypergraph representation for WSIs is another important contribution of this thesis, offering a way to represent complex tissue architecture and spatial relationships while maintaining biological interpretability.

Several key explainability features highlight the value of the hypergraph approach. First, it enables the visualization of community and spatial structures, capturing higher-order relationships between different regions of tissue. This makes it easier to understand and visualize how various areas of a tumor interact, helping pathologists and researchers focus on specific regions of interest, such as cellular clusters or areas with immune cell activity. Additionally, the hypergraph approach enhances the interpretability of pathological features by concentrating on spatial and community structures within WSIs. This provides clear explanations for how these features contribute to tissue pathology, offering a transparent understanding of how spatial patterns impact disease progression.

6.2.6 . Synthesis: The Role of Explainability in Multimodal Data Integration

The various methods developed in this thesis demonstrate a clear commitment to making complex data analysis both explainable and interpretable. By emphasizing transparency throughout the process, from the integration of molecular-level multi-omics data to tissue-level spatial analysis, these methods ensure that clinicians and researchers alike can access and understand the results.

The contributions to explainability can be seen in several key areas. First, these methods serve to bridge the gap between computational models and clinical expertise, ensuring that the sophisticated models used to analyze multimodal data are comprehensible to clinical experts, thereby facilitating their application in real-world decision-making. Additionally, the focus on explainability enhances trust in the outputs of machine learning models, which is crucial for their acceptance and integration into medical practice. Lastly, by making the results explainable, these methods provide actionable insights that can directly inform clinical decisions. This is particularly important in precision oncology, where a deep understanding of patient-specific features is essential for developing personalized treatment strategies.

The contributions of this thesis to the field of explainability are diverse and impactful. Methodologically, the development of tools such as H&Explainer and GMM-CeFlow represents a significant advancement in computational pathology, ensuring that the complex analyses of high-dimensional data are transparent and easy to interpret. Clinically, these methods offer the potential to improve cancer diagnostics, prognostics, and treatment planning by providing explainable and actionable in-

sights from multimodal data, ultimately enhancing patient outcomes.

6.3 . Challenges & Opportunities for Multimodal Integration

Integrating bulk omics data, such as genomics or transcriptomics, with histopathology slides presents several significant challenges. One of the primary limitations stems from the inherent differences in data localization between these modalities. Bulk omics data typically provide averaged molecular profiles across a heterogeneous mixture of cells, offering a global view of the biological processes within a sample. However, this averaging effect obscures the spatial context and cellular heterogeneity, making it difficult to relate specific molecular signals to localized tissue structures observed in histopathology slides.

Histopathology slides, on the other hand, offer high-resolution visual representations of tissue architecture, revealing the spatial organization of cells, tissue morphology, and microenvironmental context. While these images provide rich information about the structural and morphological aspects of the tissue, they need the detailed molecular characterization provided by omics data. The challenge, therefore, lies in bridging the gap between these two modalities—one that captures global molecular profiles without spatial resolution and the other that offers detailed spatial information without the associated molecular data.

The lack of localization in bulk omics data limits the ability to accurately map molecular changes to specific regions or cell types within the tissue. This disconnect can lead to difficulties in interpreting the biological significance of the integrated data, particularly when trying to understand complex interactions between different cell populations or microenvironmental factors that contribute to disease progression. For instance, the molecular signatures derived from bulk omics might not correspond directly to any single histological region, making it challenging to draw meaningful conclusions from the multimodal integration.

Emerging technologies like spatial transcriptomics have been developed to overcome these limitations. Spatial transcriptomics enables the measurement of gene expression across tissue sections while preserving spatial context, effectively combining the strengths of omics data and histopathology. This technology allows for the localization of gene expression patterns within specific regions of a tissue section, providing a more granular view of the molecular landscape about the tissue's structural features.

6.4 . Spatial Data for Precision Oncology

Spatial omics, particularly spatial transcriptomics, has emerged as a transformative technology in precision oncology, enabling the simultaneous capture of molecular and spatial data at the single-cell resolution. By integrating spatial information with traditional high-throughput molecular profiling (such as genomics, transcriptomics, proteomics, and metabolomics), this approach provides a more comprehensive understanding of tumor biology. When combined with whole-slide imaging (WSI) and bulk omics data, spatial omics offers novel insights into tissue architecture, tumor heterogeneity, and microenvironment interactions, which are critical for advancing personalized medicine in oncology [168, 178].

One of the key advantages of spatial transcriptomics is its ability to retain the spatial context of tissue samples, which is often lost in traditional bulk omics approaches. This spatial resolution is crucial for studying the intricate organization of tumors, where different cell populations and molecular features are distributed across diverse microenvironments. In combination with WSI, spatial transcriptomics enables researchers to map molecular data directly onto tissue morphology, creating a multimodal view of the tissue that links molecular alterations to specific histopathological features [229].

For instance, spatial transcriptomics can identify which regions of a tumor exhibit specific gene expression patterns, helping to pinpoint areas of interest such as regions of high cell proliferation, immune cell infiltration, or necrosis. When this spatial data is integrated with bulk omics, it provides a layered approach where bulk data offers a global molecular snapshot, while spatial transcriptomics zooms in to provide local insights, revealing the heterogeneous nature of the tumor. This combination is particularly powerful for precision oncology, as it allows clinicians and researchers to connect molecular data to specific tissue regions, guiding targeted therapeutic strategies and improving prognosis [13].

6.5 . Applications in WSI and Histopathology

Whole-slide imaging (WSI) provides high-resolution visual representations of tissue architecture, capturing critical information about the spatial organization of cells, tissue morphology, and the microenvironmental context. While WSIs offer detailed structural information, they need the molecular

characterization provided by omics data. This is where spatial transcriptomics steps in, bridging the gap by mapping gene expression profiles onto the visual features seen in histopathology [204].

Researchers can better understand the spatial relationships between cellular phenotypes and molecular states by combining WSI and spatial transcriptomics. For example, in a tumor sample, spatial transcriptomics can help delineate regions of active immune response by identifying areas where immune-related gene expression is heightened. At the same time, WSIs can provide complementary insights into the structural features of these regions, such as immune cell clustering or infiltration into the tumor. This multi-layered analysis helps identify spatial patterns critical for understanding disease progression and making more informed treatment decisions [85].

Moreover, the study of niches—spatially distinct cellular microenvironments in tissues—has been significantly advanced through spatial transcriptomics. Transposing niche detection from spatial transcriptomics to histopathology allows for the detailed characterization of molecular heterogeneity within tissue architecture. The collaboration on Novae with Quentin Blampey arose from this need to strengthen niche detection, leveraging graph-based models to identify and model spatial domains in ST data accurately. By encoding spatial and molecular relationships, Novae enhances our ability to detect and map cellular niches. These advancements can be transposed back into histopathology, where the combination of Novae's insights with WSI along with the human interpretable features presented in chapter 4 can refine the identification of tissue niches, providing a more nuanced understanding of disease progression.

This integration also supports the development of new diagnostic tools. Correlating molecular profiles with histopathological features and spatial transcriptomics, augmented by niche detection models like Novae, could enhance the accuracy of automated pathology systems. This allows for more precise segmentation of tumor regions, identification of specific cell types, and prediction of patient outcomes based on molecular and spatial data [36].

6.6 . Enhancing Bulk Omics with Spatial Data

Traditional bulk omics analyses provide an averaged molecular profile from a mixture of cells within a sample. While valuable for understanding general molecular trends, bulk data often obscures a tissue's spatial and cellular heterogeneity, making it difficult to distinguish which cells or regions are driving specific molecular changes. This limitation becomes particularly significant in oncology, where

tumors are highly heterogeneous, with different regions exhibiting distinct genetic and phenotypic profiles [237].

By integrating bulk omics data with spatial transcriptomics, researchers can overcome this limitation and disaggregate the bulk molecular data into more meaningful, spatially resolved components. Spatial transcriptomics identifies spatially distinct molecular subpopulations within the tumor, enabling researchers to pinpoint where specific molecular signals originate within the tissue architecture. This is especially useful for understanding how different cell populations (e.g., cancer cells, immune cells, stromal cells) interact within the tumor microenvironment [233].

Moreover, integrating spatial transcriptomics with bulk multi-omics data—such as genomics, proteomics, and metabolomics provides a more holistic view of the biological processes within a tumor. For example, bulk proteomic data may reveal dysregulation in signaling pathways. At the same time, spatial transcriptomics can localize these changes to specific cell types or regions within the tumor, offering more profound insights into the molecular drivers of cancer [30]. This multi-modal approach is critical for understanding tumor heterogeneity, identifying novel therapeutic targets, and improving treatment strategies [134].

6.7 . Future Directions

While the integration of spatial transcriptomics with whole-slide imaging (WSI) and bulk omics offers immense potential for precision oncology, several significant challenges must be addressed to fully harness the power of these technologies. The future of multimodal integration lies not only in refining existing approaches but also in overcoming critical obstacles such as computational complexity, data harmonization, model generalization, and handling missing data. Addressing these issues will be essential to drive innovation in both research and clinical applications.

6.7.1 . Computational Complexity and Scalability

One of the most pressing challenges in multimodal integration is the computational complexity involved in analyzing high-dimensional spatial data alongside other omics modalities. Spatial transcriptomics generates vast datasets, combining molecular and spatial information at the single-cell level. When this is integrated with bulk omics data and WSIs, the data volume increases dramatically, requiring sophisticated computational frameworks that can scale to meet these demands.

Current methods for multimodal integration often struggle to process these large and diverse datasets efficiently. This issue is exacerbated by the need to correct batch effects, align data from different modalities, and ensure that biological meaning is preserved during the integration process. Developing more scalable algorithms and frameworks capable of handling these computational challenges is essential. These tools must address efficient data storage and retrieval, particularly given the high dimensionality of spatial transcriptomics and omics data. Additionally, incorporating parallelized or distributed computing techniques is necessary to reduce processing times, making real-time clinical applications more feasible. Moreover, advanced machine learning and deep learning models, including foundation models, are required to integrate multiple data types without compromising the biological signal or predictive power.

Exploring the use of graph-based models and tensor decomposition techniques for multimodal integration could also help reduce the computational burden while maintaining high accuracy in detecting complex interactions between data layers.

6.7.2 . Standardization and Data Harmonization

As spatial transcriptomics technologies continue to advance, a significant hurdle remains the lack of standardization across platforms. Different technologies, such as 10x Genomics Visium and NanoString GeoMx, produce data with varying spatial resolutions, gene panel sizes, and technical characteristics. These differences make it difficult to integrate data across studies and platforms, thereby complicating large-scale comparative analysis and clinical translation.

To overcome this challenge, the research community will need to focus on several key areas. First, developing harmonization techniques will be essential to align data from different platforms, ensuring consistency in the processing and integration of spatial transcriptomics data with other modalities. Additionally, establishing standard operating procedures (SOPs) for data collection, processing, and analysis will be crucial to ensuring reproducibility and comparability across different studies. Achieving consensus on data formats and metadata requirements will also facilitate more seamless sharing and collaboration among research groups and institutions.

Furthermore, the creation of open-source tools and repositories for spatial omics data would support broader adoption of best practices and allow the research community to take full advantage of large, well-curated multimodal datasets.

6.7.3 . Generalization of Foundation Models

Foundation models—large-scale, pre-trained models that learn from vast, diverse datasets—have significant transformative potential for precision medicine, especially in the integration of multimodal data. These models can capture complex relationships across various biological layers, such as genomic, transcriptomic, proteomic, and spatial data, to provide highly accurate predictions regarding disease progression, treatment responses, and patient outcomes.

However, a key limitation of foundation models is their lack of generalizability when trained on datasets that are not sufficiently diverse or representative. In healthcare, where patient populations differ greatly in terms of genetics, demographics, and environmental factors, this limitation can result in biased models that may not perform well across all clinical scenarios.

Addressing this issue requires a concerted effort in several areas. First, expanding training datasets to include more diverse patient populations, disease types, and clinical settings is essential. This will help ensure that foundation models generalize well across a wide range of biological contexts, making them applicable to a broader spectrum of patients. Additionally, techniques for transfer learning need to be developed, allowing foundation models to be fine-tuned on smaller, more specific datasets without sacrificing the general knowledge acquired from large-scale training. Finally, building robust evaluation frameworks is critical to assess the performance of foundation models across different clinical settings, patient populations, and diseases. These frameworks should also incorporate fairness metrics to ensure that predictions are equitable across diverse demographic groups.

6.7.4 . Handling Missing Modalities

A common issue in multimodal integration is the presence of missing data, as not all patients may have complete datasets for every modality (e.g., genomic, transcriptomic) due to technical limitations, cost, or clinical circumstances. This challenge affects foundation models and machine learning techniques that rely on multiple data types for accurate predictions.

To enhance the robustness of multimodal integration in real-world applications, future efforts must concentrate on several key areas. First, developing advanced imputation techniques is essential, allowing models to infer missing data based on available modalities. For instance, deep learning models could predict absent transcriptomic or genomic information using spatial or histopathological data. Additionally, creating flexible models capable of dynamically adapting to the available modalities

is critical. Such models should deliver accurate predictions even when certain data types are missing, possibly through hierarchical models that prioritize specific data modalities or ensemble learning techniques that combine predictions from different data subsets. Furthermore, incorporating uncertainty quantification methods will be vital to indicate when predictions may be less reliable due to missing data. This will help clinicians better interpret model outputs and make informed decisions, even in the absence of certain modalities.

6.7.5 . Ethical Considerations and Data Privacy

As the integration of multimodal data becomes more widespread in clinical practice, ethical considerations surrounding data privacy, ownership, and the use of sensitive patient information will take on increasing importance. The detailed insights provided by spatial transcriptomics and omics data, often at the single-cell level, raise concerns about how this highly sensitive information is stored, shared, and used.

Future research and clinical applications must address several key aspects. First, there is a need to establish robust data privacy frameworks that ensure patient confidentiality while still facilitating the sharing of valuable multimodal datasets for research purposes. In addition, it is essential to ensure the ethical use of data in AI-driven healthcare systems, preventing the exploitation of patient information for commercial gain without proper consent and safeguards. Finally, transparency and explainability in the use of these advanced models must be prioritized, ensuring that both patients and clinicians have a clear understanding of how predictions are generated and can trust the outputs of these systems.

6.8 . Final Words

In this thesis, we have explored the complexities and opportunities presented by the integration of multi-omics and histopathological data in the field of precision oncology. By developing novel frameworks such as CustOmics and the hypergraph-based representation for WSIs, we have taken significant steps toward addressing the challenges posed by high-dimensional data and the intricate spatial relationships within tumors. These contributions not only enhance the ability to analyze and interpret complex biological datasets but also provide a foundation for more personalized and effective cancer treatments.

The integration of multi-modal data has proven to be a powerful approach for gaining a deeper understanding of cancer biology, as demonstrated by the innovative methodologies introduced in this work. From capturing biologically meaningful relationships across various omics layers to improving the interpretability of whole-slide images, these methods bridge the gap between computational models and clinical practice, ensuring that advanced technologies are both accessible and actionable in real-world healthcare settings.

As precision medicine continues to evolve, the frameworks and tools developed in this thesis lay the groundwork for future advancements in cancer research. However, significant challenges remain, particularly in terms of scalability, data diversity, and the ethical use of sensitive patient information. Addressing these challenges will be essential as we move toward a more integrated, data-driven approach to oncology.

In closing, this work represents a meaningful contribution to the field of precision oncology, advancing both the science of multimodal data integration and its practical application in personalized healthcare. It is my hope that the innovations presented here will inspire further research, leading to improved outcomes for cancer patients worldwide and contributing to the ongoing evolution of precision medicine.

Bibliography

[1] Hervé Abdi et al. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(2):149–179, 2013.

[2] Mohammed Adnan, Shivam Kalra, and Hamid R Tizhoosh. Representation learning of histopathology images using graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 988–989, 2020.

[3] F. Aeffner, M. D. Zarella, M. Pensky, J. A. van der Laak, M. M. Bui, and V. N. Vemuri. Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. *Journal of Pathology Informatics*, 8, 2017.

[4] Erez Lieberman Aiden et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[5] Neha Akhoon. Precision medicine: a new paradigm in therapeutics. *International journal of preventive medicine*, 12(1):12, 2021.

[6] Suhani H Almal and Harish Padh. Implications of gene copy-number variation in health and diseases. *Journal of human genetics*, 57(1):6–13, 2012.

[7] Óscar Álvarez-Machancoses and Juan Luis Fernández-Martínez. Using artificial intelligence methods to speed up drug discovery. *Expert opinion on drug discovery*, 14(8):769–777, 2019.

[8] Deepak Anand, Shrey Gadiya, and Amit Sethi. Histographs: graphs in histopathology. In *Medical Imaging 2020: Digital Pathology*, volume 11320, pages 150–155. SPIE, 2020.

[9] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

[10] Daniel A Arber, Attilio Orazi, Robert Hasserjian, and et al. The 2016 revision to the world health organization classification of myeloid neoplasms and acute leukemia. *Blood*, 127(20):2391–2405, 2016.

[11] Elisabetta Argenzio and Daniel Klimmeck. On the molecular, cellular and tissue origin of cancer, 2021.

[12] R. Arriagada, B. Bergman, A. Dunant, T. Le Chevalier, J.P. Pignon, and J. Vansteenkiste. Cisplatin-based adjuvant chemotherapy in patients with completely resected non-small-cell lung cancer. *New England Journal of Medicine*, 350(4):351–360, 2004.

[13] Michaela Asp, Joseph Bergenstråhle, and Joakim Lundeberg. Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays*, 42(10):e1900221, 2020.

[14] et al. Author1. Si-mil: Taming deep mil for self-interpretability in gigapixel histopathology. *arXiv preprint arXiv:2312.15010*, 2023.

[15] Shekoofeh Azizi et al. Big self-supervised models advance medical image classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3478–3488, 2021.

[16] Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021.

[17] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2019.

[18] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.

[19] David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, 2004.

[20] Stephen B. Baylin et al. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? *Nature Reviews Cancer*, 11(7):410–420, 2011.

[21] Stephen B. Baylin and Peter A. Jones. Dna methylation: a multistep process in human cancer. *Nature Reviews Cancer*, 6(7):503–514, 2006.

[22] Rafael Bejar, Kristin E Stevenson, Barbara A Caughey, and et al. Validation of a prognostic model and the impact of mutations in patients with lower-risk myelodysplastic syndromes. *Journal of Clinical Oncology*, 29(24):3376–3384, 2011.

[23] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[24] Hakim Benkirane, Yoann Pradat, Stefan Michiels, and Paul-Henry Cournède. Customics: A versatile deep-learning based strategy for multi-omics integration. *PLoS Computational Biology*, 19(3):e1010921, 2023.

[25] Axel Benner, Manuela Zucknick, Thomas Hielscher, Carina Ittrich, and Ulrich Mansmann. High-dimensional cox models: the choice of penalty as part of the model building process. *Biometrical Journal*, 52(1):50–69, 2010.

[26] Elsa Bernard, Heinrich Tuechler, Peter L Greenberg, Robert P Hasserjian, Javier E Arango, Yasuhito Nannya, and et al. Molecular international prognostic scoring system for myelodysplastic syndromes. *NEJM Evidence*, 1:1–14, 2022.

[27] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, and Luciano Milanesi. Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinformatics*, 17(2):S15, 2016.

[28] Marina Bibikova et al. High density dna methylation array with single cpg site resolution. *Genomics*, 89(3):384–389, 2007.

[29] Christopher M Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:645–678, 2006.

[30] Bernd Bodenmiller. Multiplexed epitope-based tissue imaging for discovery and healthcare applications. *Cell Systems*, 2(4):225–238, 2016.

[31] Francesco Bodria, Riccardo Guidotti, Fosca Giannotti, and Dino Pedreschi. Transparent latent space counterfactual explanations for tabular data. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE, 2022.

[32] Eric Bonnet, Laurence Calzone, and Tom Michoel. Integrative multi-omics module network inference with lemon-tree. *PLoS computational biology*, 11(2):e1003983, 2014.

[33] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[34] B. Brown et al. Representation learning of histopathology images using graph neural networks. *IEEE Explore*, 34:5678–5689, 2023.

[35] Wouter Bulten et al. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*, 21(2):233–241, 2020.

[36] David J Burgess. Spatial transcriptomics coming of age. *Nature Reviews Genetics*, 20(8):431, 2019.

[37] William S Bush, Scott M Dudek, and Marylyn D Ritchie. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. In *Biocomputing 2009*. World Scientific, 2009.

[38] Gabriel Campanella et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.

[39] Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anaïs Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, 12(1):124, 2021.

[40] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. *arXiv preprint arXiv:1807.05520*, 2018.

[41] Ethan Cerami et al. cbioportal for cancer genomics. *cBioPortal*, 2023.

[42] Navdeep S Chandel. Metabolism of proliferating cells. *Cold Spring Harbor Perspectives in Biology*, 13(10):a040618, 2021.

[43] Martin Charachon, Paul-Henry Cournede, Céline Hudelot, and Roberto Ardon. Leveraging conditional generative models in a general explanation framework of classifier decisions. *Future Generation Computer Systems*, 132:223–238, 2022.

[44] Kumardeep Chaudhary, Olivier Bertrand Poirion, Liangqun Lu, and Lana Garmire. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, 24(6):1248–59, 2017.

[45] Gul-e-Saba Chaudhry, Abdah Md Akim, Yeong Yik Sung, and Tengku Muhammad Tengku Sifzizul. Cancer and apoptosis: The apoptotic activity of plant and marine natural products and their potential as targeted cancer therapeutics. *Frontiers in Pharmacology*, 13:842376, 2022.

[46] Anika Cheerla and Olivier Gevaert. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, 35(14):i446–i454, 2019.

[47] Huifang Chen, Ola Engkvist, Yinhai Wang, et al. The rise of deep learning in drug discovery. *Drug Discovery Today*, 23(6):1241–1250, 2018.

[48] Kaitao Chen, Shiliang Sun, and Jing Zhao. Camil: Causal multiple instance learning for whole slide image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1120–1128, 2024.

[49] Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–349. Springer, 2021.

[50] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025, 2021.

[51] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022.

[52] Shuang Chen, Hui Xu, Chunguang Guo, Zaoqu Liu, and Xinwei Han. The role of multi-omics variants in tumor immunity and immunotherapy. *Frontiers in Immunology*, 13:1098825, 2022.

[53] Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14(4):e1006076, 2018.

[54] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[55] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.

[56] Ozan Ciga, Tony Xu, Sharon Nofech-Mozes, Shawna Noy, Fang-I Lu, and Anne L Martel. Overcoming the limitations of patch-based learning to detect cancer in whole slide images. *Scientific Reports*, 11(1):1–10, 2021.

[57] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyo, ..., and A. Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, 2018.

[58] Pierre Courtiol, Carine Maussion, Mehdi Moarii, Estelle Pronier, Stéphane Pilcer, Mohamed Sefta, Gérald Manceau, Thierry Clozel, Jonathan Wicinski, Laurent Lanta, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature Medicine*, 25(10):1519–1525, 2019.

[59] Lisa M. Coussens and Zena Werb. Cancer-related inflammation. *Nature*, 420:860–867, 2011.

[60] David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[61] Rodrigo GB Cruz, Stephen F Madden, Kieran Brennan, and Ann M Hopkins. A transcriptional link between her2, jam-a and foxa1 in breast cancer. *Cells*, 11(4):735, 2022.

[62] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94, 2019.

[63] H. Joachim Deeg and Brenda M Sandmaier. Who is fit for allogeneic transplantation? *Blood*, 130(8):830–837, 2017.

[64] C. Denkert, G. von Minckwitz, S. Darb-Esfahani, B. I. Lederer, B. I. Heppner, K. E. Weber, ..., and S. Loibl. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *The Lancet Oncology*, 19(1):40–50, 2018.

[65] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[66] Donglin Di, Shengrui Li, Jun Zhang, and Yue Gao. Ranking-based survival prediction on histopathological whole-slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 428–438. Springer, 2020.

[67] Donglin Di, Jun Zhang, Fuqiang Lei, Qi Tian, and Yue Gao. Big-hypergraph factorization neural network for survival prediction from whole slide image. *IEEE Transactions on Image Processing*, 31:1149–1160, 2022.

[68] Nikolaos Dimitriou, Ognjen Arandjelović, and Peter D Caie. Artificial intelligence in digital pathology for the classification of cancer tissue samples. *Computational and Structural Biotechnology Journal*, 17:447–456, 2019.

[69] J. Doe et al. Graph neural network for representation learning of lung cancer. *BMC Cancer*, 23:1234–1245, 2023.

[70] Achim Dolnik, Julia C Engelmann, Martina Scharfenberger-Schmeer, and et al. Clonal evolution of acute myeloid leukemia with flt3-itd mutation under treatment with the tyrosine kinase inhibitor midostaurin. *Blood*, 131(24):2632–2643, 2018.

[71] Khalid El Bairi, Harry R Haynes, Elizabeth Blackley, Susan Fineberg, Jeffrey Shear, Sophia Turner, Juliana Ribeiro De Freitas, Daniel Sur, Luis Claudio Amendola, Masoumeh Gharib, et al. The tale

of tils in breast cancer: a report from the international immuno-oncology biomarker working group. *NPJ Breast Cancer*, 7(1):150, 2021.

[72] Susan Elmore. Apoptosis: a review of programmed cell death. *Toxicologic Pathology*, 35(4):495–516, 2007.

[73] Manel Esteller. Dna methylation and cancer therapy: new developments and expectations. *Current Opinion in Oncology*, 19(1):68–75, 2007.

[74] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, ..., and J. Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019.

[75] Andre Esteva, Brett Kuprel, Roberto A. Novoa, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

[76] Navid Farahani et al. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, 7:23–33, 2015.

[77] Jawad Fares, Mohamad Y Fares, Hussein H Khachfe, Hamza A Salhab, and Youssef Fares. Molecular principles of metastasis: a hallmark of cancer revisited. *Signal transduction and targeted therapy*, 5(1):28, 2020.

[78] Andrew P. Feinberg. The epigenetic basis of common human disease. *Nature*, 447(7143):433–440, 2010.

[79] Pierre Fenaux, Ghulam J Mufti, Eva Hellström-Lindberg, and et al. Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase iii study. *The Lancet Oncology*, 10(3):223–232, 2009.

[80] Kaleigh Fernald and Manabu Kurokawa. Evading apoptosis in cancer. *Trends in cell biology*, 23(12):620–633, 2013.

[81] Andrew H. Fischer et al. Hematoxylin and eosin staining of tissue and cell sections. *Cold Spring Harbor Protocols*, 3(1):pdb–prot4986, 2008.

[82] Alex Folch-Fortuny et al. Using kernel-pca for nonlinear feature extraction in time-series prediction problems. *Neurocomputing*, 123:183–194, 2017.

[83] David A Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.

[84] Wolf H. Fridman et al. The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews Cancer*, 12:298–306, 2012.

[85] Yuqi Fu, Maura Esposito, Benjamin Boëda, Miriam Tarbier, Daniel Hornburg, Valentina Cianfanelli, Koen M O Galenkamp, Stefan J van Heeringen, Max D Wellenstein, and Andrea Sacchetti. Spatially resolved transcriptomics reveals gene expression patterns along tissue axes. *Nature Communications*, 13(1):1–14, 2022.

[86] Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *International Conference on Machine Learning*, pages 3419–3430. PMLR, 2020.

[87] Levi A. Garraway and Eric S. Lander. Genomics-driven oncology: framework for an emerging paradigm. *The New England Journal of Medicine*, 369(2):175–187, 2013.

[88] Ramiro Garzon, George A Calin, and Carlo M Croce. Micrornas in cancer. *Annual Review of Medicine*, 60:167–179, 2010.

[89] Devottam Gaurav and Sanju Tiwari. Interpretability vs explainability: The black box of machine learning. In *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, pages 523–528. IEEE, 2023.

[90] Moritz Gerstung et al. The evolutionary history of 2,658 cancers. *Nature*, 578:122–128, 2020.

[91] Peter Gibson et al. Antigen retrieval for immunohistochemistry: do we know what we are doing yet? *Journal of Histotechnology*, 29(2):99–109, 2006.

[92] Fabien Girka, Etienne Camenen, Caroline Peltier, Arnaud Gloaguen, Vincent Guillemot, Laurent Le Brusquet, and Arthur Tenenhaus. Multiblock data analysis with the rgcca package. *Journal of Statistical Software*, pages 1–36, 2023.

[93] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[94] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[95] Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18 17-18:2529–45, 1999.

[96] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019.

[97] Peter L Greenberg, Heinz Tuechler, Julie Schanz, and et al. Revised international prognostic scoring system for myelodysplastic syndromes. *Blood*, 120(12):2454–2465, 2012.

[98] John Griffin and Darren Treanor. Digital pathology: current status and future directions. *Annual Review of Pathology: Mechanisms of Disease*, 12:305–328, 2020.

[99] Stephen Grossberg. Recurrent neural networks. *Scholarpedia*, 8(2):1888, 2013.

[100] Robert L. Grossman, Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–12, 2016.

[101] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018.

[102] Qirui Guan et al. Histopathological image analysis using deep learning: opportunities and challenges. *IEEE Reviews in Biomedical Engineering*, 12:56–73, 2019.

[103] Claudia Haferlach, Yoshimitsu Nagata, Valentin Grossmann, Yusuke Okuno, Ulrike Bacher, Go Nagae, Susanne Schnittger, Masashi Sanada, H Phillip Koeffler, Lynn Y Shih, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia*, 28(2):241–247, 2014.

[104] Claudia Haferlach, Melanie Zenger, Eric Lécuyer, and et al. Molecular genetic characterization of the myelodysplastic syndrome subtype mds-u. *American Journal of Hematology*, 89(8):743–749, 2014.

[105] William Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, 2017.

[106] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[107] David J Hand and Robert J Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45:171–186, 2001.

[108] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics*, 14:1–15, 2013.

[109] Frank E. Harrell, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

[110] Jr Harrell, Frank E., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247:2543–2546, 1982.

[111] Reuben S Harris. Cancer mutation signatures, dna damage mechanisms, and potential clinical implications. *Genome medicine*, 5:1–3, 2013.

[112] Sahar Hassani, Mohamed Hanafi, El Mostapha Qannari, and Achim Kohler. Deflation strategies for multi-block principal component analysis revisited. *Chemometrics and Intelligent Laboratory Systems*, 120:154–68, 2013.

[113] Seyed Hossein Hassanpour and Mohammadamin Dehghani. Review of cancer from perspective of molecular. *Journal of cancer research and practice*, 4(4):127–129, 2017.

[114] Lin He and Gregory J Hannon. Micrornas join the p53 network—another piece in the tumour-suppression puzzle. *Nature Reviews Cancer*, 4(10):767–774, 2004.

[115] Irina Higgins, Loic Matthey, Arka Pal, et al. beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR)*, 2017.

[116] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[117] Muta Hira, Mohammad Abdur Razzaque, Claudio Angione, James Scrivens, Saladin Sawan, and Mosharraf Sarkar. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Scientific Reports*, 11(1):6265, 2021.

[118] Zhi Huang, Xiaohui Zhan, Shunian Xiang, Travis S. Johnson, Bryan Helm, Christina Y. Yu, Jie Zhang, Paul Salama, Maher Rizkalla, Zhi Han, and Kun Huang. Salmon: Survival analysis learning with multi-omics neural networks on breast cancer. *Frontiers in Genetics*, 10:166, 2019.

[119] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, and Jens Kleesiek. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024.

[120] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[121] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, page 448–456. PMLR, 2015.

[122] Jeremy R. Jass. Significance of inflammatory infiltrates in colorectal cancer. *Histopathology*, 40(1):50–55, 2001.

[123] Guillaume Jaume, Anurag Vaidya, Richard Chen, Drew Williamson, Paul Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[124] Fei Jiang et al. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2(4):230–243, 2017.

[125] Long Jing and Xiaojie Tian. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Neural Networks and Learning Systems*, 31(3):871–890, 2020.

[126] Kevin B Johnson, Wei-Qi Wei, Dilhan Weeraratne, Mark E Frisse, Karl Misulis, Kyu Rhee, Juan Zhao, and Jane L Snowdon. Precision medicine, ai, and the future of personalized health care. *Clinical and translational science*, 14(1):86–93, 2021.

[127] Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[128] Peter A. Jones and Stephen B. Baylin. The role of dna methylation in cancer. *Cell*, 128(5):683–692, 2007.

[129] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

[130] Mrinalini R. Junttila and Gerard I. Evan. Effects of necrotic cells on cancer therapy. *Nature Reviews Cancer*, 13:801–808, 2013.

[131] Min-Gu Kang, Hye-Ran Kim, Bo-Young Seo, Jun Hyung Lee, Seok-Yong Choi, Soo-Hyun Kim, Jong-Hee Shin, Soon-Pal Suh, Jae-Sook Ahn, and Myung-Geun Shin. The prognostic impact of mutations in spliceosomal genes for myelodysplastic syndrome patients without ring sideroblasts. *BMC cancer*, 15:1–11, 2015.

[132] Edward L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[133] Jared Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.

[134] Ruichen Ke et al. In situ sequencing for rna analysis in preserved tissue and cells. *Nature Methods*, 10(9):857–860, 2013.

[135] Hyun J. Kim et al. Immunohistochemistry for pathologists: protocols, pitfalls, and tips. *Journal of Pathology and Translational Medicine*, 40(5):357–371, 2006.

[136] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[137] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.

[138] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.

[139] Vasiliki Koliaraki et al. Fibroblasts in inflammation and tissue repair: versatile and dynamic players in disease pathogenesis. *Current Opinion in Pharmacology*, 55:72–79, 2020.

[140] Daisuke Komura and Sozo Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018.

[141] Konstantina Kourou et al. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

[142] David B. Krizman. Digital pathology: A practical guide. *Archives of Pathology & Laboratory Medicine*, 139(12):1715–1721, 2015.

[143] Peter W. Laird. Principles and challenges of genome-wide dna methylation analysis. *Nature Reviews Genetics*, 11(3):191–203, 2010.

[144] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[145] Anna Maria Lewandowska, Marcin Rudzki, Sławomir Rudzki, Tomasz Lewandowski, and Barbara Laskowska. Environmental risk factors for cancer-review paper. *Annals of Agricultural and Environmental Medicine*, 26(1):1–7, 2018.

[146] Feng Li, Tan Wu, Yanjun Xu, Qun Dong, Jing Xiao, Yingqi Xu, Qian Li, Chunlong Zhang, Jianxia Gao, Liqiu Liu, et al. A comprehensive overview of oncogenic pathways in human cancer. *Briefings in bioinformatics*, 21(3):957–969, 2020.

[147] Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 174–182. Springer, 2018.

[148] Zheng Li et al. Artificial intelligence in cancer therapy. *Artificial Intelligence in Medicine*, 98:1–6, 2019.

[149] Arthur Liberzon, Caron Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database (msigdb) hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015.

[150] Huangjing Lin, Hao Chen, Simon Graham, Qi Dou, Nasir Rajpoot, and Pheng-Ann Heng. Fast scannet: Fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection. *IEEE transactions on medical imaging*, 38(8):1948–1958, 2019.

[151] Weiping Lin, Zhenfeng Zhuang, Lequan Yu, and Liansheng Wang. Boosting multiple instance learning models for whole slide image classification: A model-agnostic framework based on counterfactual inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[152] Alan List, Gordon Dewald, John Bennett, Aristotle Giagounidis, Azra Raza, Eric Feldman, Bayard Powell, Peter Greenberg, Deborah Thomas, Richard Stone, et al. Lenalidomide in the myelodysplastic syndrome with chromosome 5q deletion. *New England Journal of Medicine*, 355(14):1456–1465, 2006.

[153] Geert Litjens et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[154] Chengming Liu, Sufei Zheng, Runsen Jin, Xinfeng Wang, Feng Wang, Ruochuan Zang, Haiyan Xu, Zhiliang Lu, Jianbing Huang, Yuanyuan Lei, et al. The superior efficacy of anti-pd-1/pd-l1 immunotherapy in kras-mutant non-small cell lung cancer that correlates with an inflammatory phenotype and increased immunogenicity. *Cancer letters*, 470:95–105, 2020.

[155] Jian Liu et al. Multi-omics integration with weighted affinity and self-diffusion applied for cancer subtypes identification. *Journal of Translational Medicine*, 2023.

[156] S Loi, S Michiels, S Adams, S Loibl, J Budczies, C Denkert, and R Salgado. The journey of tumor-infiltrating lymphocytes as a biomarker in breast cancer: clinical utility in an era of checkpoint inhibition. *Annals of Oncology*, 32(10):1236–1244, 2021.

[157] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[158] Mingyang Lu, Jie Wu, Xueyu Dong, Xia Wang, Xin Han, Yanan Zhang, Chuanjian Lu, and Chenglong Wu. Deep learning-based integration of histopathological images and genomic data predicts lung cancer recurrence. *Nature Communications*, 12(1):1–11, 2021.

[159] Wenqi Lu, Michael Toss, Muhammad Dawood, Emad Rakha, Nasir Rajpoot, and Fayyaz Minhas. Slidegraph+: Whole slide image level graphs to predict her2 status in breast cancer. *Medical Image Analysis*, page 102486, 2022.

[160] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[161] Brendon Lutnick, David Manthey, and Pinaki Sarder. A tool for user friendly, cloud based, whole slide image segmentation. *arXiv preprint arXiv:2101.07222*, 2021.

[162] Anant Madabhushi and Geunbae Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016.

[163] Mahsa Mahdizadehi, Marie Saghaeian Jazi, Seyyed Mostafa Mir, and Seyyed Mehdi Jafari. Role of fibrilins in human cancer: A narrative review. *Health Science Reports*, 6(7):e1434, 2023.

[164] Aridoss Manimaran, Dhasarathan Chandramohan, SG Shrinivas, and N Arulkumar. A comprehensive novel model for network speech anomaly detection system using deep learning approach. *International Journal of Speech Technology*, 23(2):305–313, 2020.

[165] G. Marie et al. Tissue damage and repair in periodontitis: pathophysiology and clinical implications. *Journal of Clinical Periodontology*, 28(8):681–688, 2001.

[166] Niccolò Marini, Sebastian Otálora, Damian Podareanu, Mart van Rijthoven, Jeroen van der Laak, Francesco Ciompi, Henning Müller, and Manfredo Atzori. Multi_scale_tools: a python library to exploit multi-scale whole slide images. *Frontiers in Computer Science*, 3:684521, 2021.

[167] Mireya Martínez-García and Enrique Hernández-Lemus. Data integration challenges for machine learning in precision medicine. *Frontiers in medicine*, 8:784455, 2022.

[168] Vivien Marx. A method to the madness. *Nature Methods*, 16(1):9–12, 2019.

[169] Annika McCarthy et al. Applications of machine learning in cancer. *Journal of Clinical Oncology*, 38(9):2342–2350, 2020.

[170] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3:861, 09 2018.

[171] Hassane Medyouf. The microenvironment in human myeloid malignancies: Emerging concepts and therapeutic implications. *Blood*, 129(12):1617–1626, 2017.

[172] Arjun Mehta and Debu Tripathy. Co-targeting estrogen receptor and her2 pathways in breast cancer. *The breast*, 23(1):2–9, 2014.

[173] Xu Min, Xiaojie Tian, Jinmei Liu, et al. Survey of clustering algorithms. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):4316–4333, 2018.

[174] Biswapriya Misra, Carl Langefeld, Michael Olivier, and Laura Cox. Integrated omics: Tools, advances, and future approaches. *Journal of Molecular Endocrinology*, 61(1):R21–45, 2018.

[175] P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. Vega, ..., and L. A. Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

[176] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

[177] Andrei Moga et al. Machine learning for histopathological images analysis: Major trends and challenges. *Computational and Structural Biotechnology Journal*, 17:177–190, 2019.

[178] Robert Moncada et al. Integrating multi-omics and spatial data to chart tumor ecosystems. *Nature Reviews Cancer*, 20(5):271–286, 2020.

[179] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, 5(7):621–628, 2008.

[180] F Mosele, J Remon, J Mateo, CB Westphalen, Fabrice Barlesi, MP Lolkema, N Normanno, A Scarpa, Mark Robson, F Meric-Bernstam, et al. Recommendations for the use of next-generation sequencing (ngs) for patients with metastatic cancers: a report from the esmo precision medicine working group. *Annals of Oncology*, 31(11):1491–1505, 2020.

[181] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

[182] Narasimhan Nagan, Shakti Ramkissoon, Marcia Eisenberg, Anjen Chenn, and Taylor J Jensen. Impact of the updated ipss-molecular prognostic scoring system for myelodysplastic syndrome in 10,283 real world samples. *Blood*, 140(Supplement 1):9783–9784, 2022.

[183] National Cancer Institute. The cancer genome atlas program (tcga). *NCI*, 2023.

[184] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[185] Kee Yuan Ngiam and IW Khor. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5):e262–e273, 2019.

[186] John S. Nielsen et al. Stromal extracellular matrix promotes the survival of mammary epithelial cells and tissue architecture. *Nature Cell Biology*, 3:697–704, 2001.

[187] Peter C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194:23–28, 1976.

[188] Jérôme Pagès. *Multiple Factor Analysis by Example Using R*. CRC Press, 2014.

[189] Liron Pantanowitz et al. Review of the current state of whole slide imaging in pathology. *Journal of Pathology Informatics*, 1(1):36, 2010.

[190] Elli Papaemmanuil, Moritz Gerstung, Luca Malcovati, and et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*, 122(22):3616–3627, 2013.

[191] Elli Papaemmanuil, Moritz Gerstung, Luca Malcovati, Serena Tauro, Gunes Gundem, Peter Van Loo, Connie J Yoon, Peter Ellis, David C Wedge, Andrea Pellagatti, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*, 122(22):3616–3627, 2013.

[192] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[193] Hector Peinado et al. Necrotic cell death and inflammation in cancer development. *Nature Reviews Cancer*, 17:104–115, 2017.

[194] João Pessoa, Marta Martins, Sandra Casimiro, Carlos Pérez-Plasencia, and Varda Shoshan-Barmatz. Altered expression of proteins in cancer: function and potential therapeutic targets. *Frontiers in Oncology*, 12:949139, 2022.

[195] Luís Pinto-Coelho. How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications. *Bioengineering*, 10(12):1435, 2023.

[196] Antoine Pirovano, Hippolyte Heuberger, Sylvain Berlemont, Saïd Ladjal, and Isabelle Bloch. Improving interpretability for computer-aided diagnosis tools on whole slide imaging with multiple instance learning and gradient-based explanations. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3*, pages 43–53. Springer, 2020.

[197] Peter Priestley et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*, 575:210–216, 2019.

[198] Darwin J. Prockop. Marrow stromal cells as stem cells for nonhematopoietic tissues. *Science*, 276(5309):71–74, 1997.

[199] Deepak Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

[200] Wayne S. Rasband. Imagej: Image processing and analysis in java. *Astrophysics Source Code Library*, page ascl:1206.013, 2012.

[201] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

[202] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[203] Dan R. Robinson et al. Integrative clinical genomics of metastatic cancer. *Nature*, 548:297–303, 2017.

[204] Alexis J. Rodriguez et al. The power of single-cell rna sequencing for spatial transcriptomics. *Nature Biotechnology*, 39(5):639–641, 2021.

[205] Juan Rosai. *Rosai and Ackerman's Surgical Pathology*. Elsevier, 2004.

[206] Jane E. Ross. *Biomedical Visualisation*. Springer, 2014.

[207] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning Internal Representations by Error Propagation*. MIT Press, 1986.

[208] Erkki Ruoslahti. Integrins as therapeutic targets. *Annals of Medicine*, 29(1):53–57, 1997.

[209] Samir Sadok, Simon Leglaive, Laurent Girin, Xavier Alameda-Pineda, and Renaud Séguier. A multimodal dynamical variational autoencoder for audiovisual speech representation learning. *Neural Networks*, 172:106120, 2024.

[210] Ofer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[211] R. Salgado, C. Denkert, S. Demaria, N. Sirtaine, F. Klauschen, G. Pruneri, ..., and S. Loi. The evaluation of tumor-infiltrating lymphocytes (tils) in breast cancer: recommendations by an international tils working group 2014. *Annals of Oncology*, 26(2):259–271, 2015.

[212] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018.

[213] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018.

[214] Stephen-John Sammut, Mireia Crispin-Ortuzar, Suet-Feung Chin, Elena Provenzano, Helen A Bardwell, Wenxin Ma, Wei Cope, Ali Dariush, Sarah-Jane Dawson, Jean E Abraham, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*, 601(7894):623–629, 2022.

[215] Alex Sánchez, José Fernández-Real, Esteban Vegas, Francesc Carmona, Jacques Amar, Remy Burcelin, Matteo Serino, Francisco Tinahones, M Carmen Ruíz de Villa, Antonio Minãrro, et al. Multivariate methods for the integration and visualization of omics data. In *Spanish Symposium on Bioinformatics*, page 382. Springer.

[216] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[217] Kornel E. Schuebel et al. The epigenome as a target for cancer therapy. *Nature Reviews Drug Discovery*, 6(7):580–591, 2007.

[218] Michael Schwalbe et al. Classical machine learning for bioinformatics. *Nature Reviews Genetics*, 20(3):117–129, 2019.

[219] Lauren M. Schwartz et al. Autophagy, senescence, and cancer: a review. *Frontiers in Bioscience*, 21:1170–1186, 2016.

[220] PJ Schüffler, TJ Fuchs, CS Ong, V Roth, and JM Buhmann. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nature Communications*, 2021.

[221] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

[222] Dezna C. Sheehan and Barbara B. Hrapchak. *Theory and practice of histotechnology*. Mosby, 2008.

[223] Dinggang Shen et al. Analysis of functional mri data by kernel canonical correlation analysis. *Human Brain Mapping*, 25(3):111–121, 2005.

[224] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, Ahmedin Jemal, et al. Cancer statistics, 2023. *Ca Cancer J Clin*, 73(1):17–48, 2023.

[225] Nikola Simidjievski, Cristian Bodnar, Ifrah Tariq, Paul Scherer, Helena Andres Terre, Zohreh Shams, Mateja Jamnik, and Pietro Liò. Variational autoencoders for cancer data integration: Design principles and computational practice. *Frontiers in Genetics*, 10:1205, 2019.

[226] A. Smith et al. Graph neural networks in cancer and oncology research: Emerging and future trends. *MDPI Cancers*, 15:5858, 2023.

[227] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, 2012.

[228] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[229] Brett T. Staahl et al. Visualize with visium: spatially resolved gene expression in tissue sections. *BioTechniques*, 58(5):286–289, 2016.

[230] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[231] PJ Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, and Paul Honeine. Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications*, 117:103–111, 2019.

[232] Noriaki Sunaga, David S Shames, Luc Girard, Michael Peyton, Jill E Larsen, Hisao Imai, Junichi Soh, Mitsuo Sato, Noriko Yanagitani, Kyoichi Kaira, et al. Knockdown of oncogenic kras in non–small cell lung cancers suppresses tumor growth and sensitizes tumor cells to targeted therapy. *Molecular cancer therapeutics*, 10(2):336–346, 2011.

[233] Valentin Svensson, Sarah A Teichmann, and Oliver Stegle. Spatialde: identification of spatially variable genes. *Nature Methods*, 15(5):343–346, 2018.

[234] Ming Tan et al. Unsupervised learning: the foundations and challenges of data clustering and anomaly detection. *ACM Computing Surveys*, 52(4):1–42, 2019.

[235] Christopher J Tape, Stephanie Ling, Maria Dimitriadi, Kelly M McMahon, Jonathan D Worboys, Hui Sun Leong, Ida C Norrie, Crispin J Miller, George Poulogiannis, Douglas A Lauffenburger, et al. Oncogenic kras regulates tumor cell signaling via stromal reciprocation. *Cell*, 165(4):910–920, 2016.

[236] Qiaoying Teng, Zhe Liu, Yuqing Song, Kai Han, and Yang Lu. A survey on the interpretability of deep learning in medical diagnosis. *Multimedia Systems*, 28(6):2335–2355, 2022.

[237] Tao Tian, Wei Zhang, and Dongxiao Wang. Graphst: Multi-resolution graph neural networks for spatial transcriptomics. *Nature Computational Biology*, 6:449–462, 2022.

[238] Suzanne L. Topalian, Charles G. Drake, and Drew M. Pardoll. Immune checkpoint blockade: a common denominator approach to cancer therapy. *Cancer Cell*, 27(4):450–461, 2016.

[239] Eric Topol. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, 2019.

[240] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

[241] Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

[242] Wilma van Eerdeweg and Geoffrey B. Snow. Surgical resection margins in oral cancer. *Journal of Clinical Pathology*, 43(8):564–569, 1990.

[243] Greg Van Houdt, Carlos Mosquera, and Gonzalo Nápoles. A review on the long short-term memory model. *Artificial Intelligence Review*, 53(8):5929–5955, 2020.

[244] Sjoerd Van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? *Advances in neural information processing systems*, 32, 2019.

[245] Charles Vaske, Stephen Benz, J Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–45, 2010.

[246] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[247] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *ICLR*, 2018.

[248] G. Viaud, P. Mayilvahanan, and P. Cournede. Representation learning for the clustering of multi-omics data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19:135–45, 2021.

[249] Pascal Vincent et al. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103. ACM, 2008.

[250] Kanchan Vishnoi, Navin Viswakarma, Ajay Rana, and Basabi Rana. Transcription factors in cancer development and therapy. *Cancers*, 12(8):2296, 2020.

[251] Fei Wang et al. Artificial intelligence in medicine: current applications and future directions. *Computers in Biology and Medicine*, 116:103595, 2019.

[252] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[253] Zixu Wang, Xin Chen, et al. Mogonet: A multi-omics integration framework using graph attention network for biomedical classification. *Frontiers in Genetics*, 12:724615, 2021.

[254] M. D. Wellenstein and K. E. de Visser. Cancer-cell-intrinsic mechanisms shaping the tumor immune landscape. *Immunity*, 48(3):399–416, 2018.

[255] Zachary Wevodau et al. Integration of incomplete multi-omics data using knowledge distillation and supervised variational autoencoders for disease progression prediction. *Journal of Biomedical Informatics*, 147:104512, 2023.

[256] Eloise Withnell, Xiaoyu Zhang, Kai Sun, and Yike Guo. Xomivae: an interpretable deep learning model for cancer classification using high-dimensional omics data. *Briefings in Bioinformatics*, 22:315, 2021.

[257] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[258] Wenzhuo Yang, Jia Li, Caiming Xiong, and Steven CH Hoi. Mace: an efficient model-agnostic framework for counterfactual explanation. *arXiv preprint arXiv:2205.15540*, 2022.

[259] Zhiyong Yang, Qianqian Xu, Shilong Bao, and Xiaochun Cao. Learning with multiclass auc: Theory and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 2021.

[260] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

[261] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

[262] Gilad Yehudai, Ethan Fetaya, Eli Meirom, Gal Chechik, and Haggai Maron. From local structures to size generalization in graph neural networks. In *International Conference on Machine Learning*, pages 11975–11986. PMLR, 2021.

[263] Gokul Yenduri, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Rutvij H Jhaveri, Weizheng Wang, Athanasios V Vasilakos, Thippa Reddy Gadekallu, et al. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *arXiv preprint arXiv:2305.10435*, 2023.

[264] Jie Yi, Liwen Ren, Dandan Li, Jie Wu, Wan Li, Guanhua Du, and Jinhua Wang. Trefoil factor 1 (tff1) is a potential prognostic biomarker with functional significance in breast cancers. *Biomedicine & Pharmacotherapy*, 124:109827, 2020.

[265] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *NeurIPS*, 2018.

[266] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017.

[267] Annie Yuan, John Haanen, and Riccardo Mezzadra. A comprehensive review of the roles of tumour-associated macrophages in tumour development and progression. *Nature Reviews Clinical Oncology*, 17:441–454, 2020.

[268] Boudewijn Zaan et al. Margins of resection in colorectal cancer. *Diseases of the Colon & Rectum*, 44(7):1021–1028, 2001.

[269] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

[270] Amer M Zeidan, Heike A Knaus, Theresa M Robinson, and et al. Immunomodulatory drugs in hematologic malignancies: Connecting immunology and hematology. *Experimental Hematology*, 71:38–49, 2019.

[271] et al. Zhang, Zizhao. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1:236–45, May 2019.

[272] Shihua Zhang, Qingjiao Li, Juan Liu, and Xianghong Zhou. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. *Bioinformatics*, 27(13):i401–09, 2011.

[273] Xiaoyu Zhang, Yuting Xing, Kai Sun, and Yike Guo. Omiembed: A unified multi-task deep learning framework for multi-omics data. *Cancers*, 13(12):3047, 2021.

[274] Xiaoyu Zhang, Jingqing Zhang, Kai Sun, Xian Yang, Chengliang Dai, and Yike Guo. Integrated multi-omics analysis using variational autoencoders: Application to pan-cancer classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, page 765–69, 2019.

[275] Huakan Zhao, Lei Wu, Guifang Yan, Yu Chen, Mingyue Zhou, Yongzhong Wu, and Yongsheng Li. Inflammation and tumor progression: signaling pathways and targeted intervention. *Signal transduction and targeted therapy*, 6(1):263, 2021.

[276] Lijun Zhao et al. Multi-modal learning for regression and time-series modelling: A survey. *arXiv preprint arXiv:2006.02464*, 2020.

[277] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

[278] Xing Zhao et al. Kernel-based support vector machine with a feature selection method for financial time series forecasting. *Expert Systems with Applications*, 84:193–205, 2016.

[279] J. Zhou, J. Lu, C. Gao, C. Zeng, X. Zhang, and X. Guo. Predictive and prognostic value of tp53 status for patients with locally advanced nasopharyngeal carcinoma treated with radiation therapy: a retrospective study. *Radiation Oncology*, 12(1):124, 2017.

[280] Yanning Zhou, Simon Graham, Navid Alemi Koohbanani, Muhammad Shaban, Pheng-Ann Heng, and Nasir Rajpoot. Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2019.

[281] Zongwei Zhou, Md Moinul Haque Siddiquee, Nima Tajbakhsh, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7340–7351, 2017.

[282] Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547. IEEE, 2016.

[283] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7234–7242, 2017.

Additional Details & Mathematical Frameworks

A.1 . Multiple Instance Learning

Multiple Instance Learning (MIL) is a variation of supervised learning where the training set is composed of labeled bags (sets) of instances, rather than individually labeled instances. In traditional supervised learning, each instance is labeled and used for training, but MIL addresses situations where labels are assigned to groups of instances (bags) without specifying which instance within the group is responsible for the label. This learning paradigm is particularly useful in domains where obtaining detailed annotations is challenging or expensive.

A.1.1 . Problem Formulation

In Multiple Instance Learning, the primary goal is to train a model using a dataset consisting of bags, each labeled as positive or negative. A bag is labeled as positive if it contains at least one positive instance, and negative if all instances within it are negative. Formally, let $X = \{X_1, X_2, \ldots, X_n\}$ be a set of bags, where each bag X_i contains instances $\{x_{i1}, x_{i2}, \ldots, x_{im_i}\}$. Each bag X_i is associated with a label $Y_i \in \{0, 1\}$. The challenge is to learn a function f that predicts the label of new bags based on the instances they contain.

MIL assumes that while the individual instances within a positive bag might not all be positive, there is at least one instance that contributes to the bag's positive label. Conversely, a negative bag has no positive instances. This assumption is crucial in many real-world applications where fine-

grained labels are impractical to obtain.

A.1.2 . Theory

The theory behind Multiple Instance Learning involves understanding how to aggregate information from instances within a bag to make predictions about the bag's label. One common approach is to use an instance-level classifier to evaluate each instance and then aggregate these evaluations to predict the bag's label. For example, a maximum function might be used, where the bag is labeled positive if any instance is predicted to be positive.

Another approach is to use a specialized MIL model, such as MIL-based support vector machines (SVM), neural networks, or ensemble methods. These models are designed to handle the ambiguity of instance labels within bags and can learn more complex relationships between instances and bag labels. For instance, the Diverse Density (DD) algorithm aims to find a region in the instance space where positive instances from positive bags are densely clustered and negative instances are sparse.

A.1.3 . Applications in Biomedical Data

Multiple Instance Learning has found significant applications in the biomedical field, particularly in histopathology. Histopathology involves the examination of tissue samples to diagnose diseases, such as cancer. Obtaining pixel-level annotations of pathological images is extremely labor-intensive and requires expert knowledge, making MIL a valuable approach.

In histopathology, a tissue sample (bag) is divided into smaller patches (instances), which are then analyzed. The entire sample might be labeled as cancerous or non-cancerous based on the presence of cancerous cells in any of the patches. MIL models can be trained to predict the overall diagnosis based on the patches without needing each patch to be individually labeled. This significantly reduces the annotation burden while still leveraging the detailed information within the tissue.

For example, in breast cancer diagnosis, whole-slide images (WSIs) of tissue samples can be divided into smaller tiles. An MIL approach can be used to classify these WSIs as benign or malignant based on the presence of malignant cells in any of the tiles. Similar applications are seen in identifying metastases in lymph nodes, grading of tumors, and detecting other pathological conditions.

The use of MIL in biomedical imaging extends beyond histopathology to areas like radiology, where similar challenges of detailed labeling exist. MIL models can assist in analyzing complex medical images by focusing on regions that contribute to the overall diagnosis, thereby improving accuracy

and efficiency in medical decision-making.

In conclusion, Multiple Instance Learning addresses the challenge of learning from weakly labeled data by focusing on the relationship between bags and their instances. Its theoretical foundations and practical applications, particularly in biomedical data and histopathology, demonstrate its potential to handle complex real-world problems where detailed annotations are scarce or costly to obtain.

A.2 . Generalities on Graph Neural Networks

Graph Neural Networks (GNNs) have emerged as a powerful tool for analyzing and learning from data structured as graphs. Unlike traditional neural networks that operate on grid-like data such as images or sequences, GNNs are designed to work with graph data, capturing the dependencies between nodes through their connections. This capability makes GNNs highly suitable for a variety of applications, including social network analysis, molecular chemistry, recommendation systems, and more. The core idea behind GNNs is to leverage the graph structure to improve learning by propagating and aggregating information across nodes and edges.

A.2.1 . Problem Formulation

A graph $G = (V, E)$ consists of a set of nodes V and a set of edges E. Each node $v_i \in V$ can have a feature vector $\mathbf{x}_i$, and each edge $(v_i, v_j) \in E$ can have an edge feature $\mathbf{e}_{ij}$. The goal of a GNN is to learn a representation $\mathbf{h}_i$ for each node v_i that captures its structural and feature information. This learned representation can be used for various tasks such as node classification, graph classification, and link prediction.

Formally, let $\mathbf{X}$ be the matrix of node features where $\mathbf{X} \in \mathbb{R}^{|V| \times d}$ and $\mathbf{A}$ be the adjacency matrix of the graph where $\mathbf{A} \in \mathbb{R}^{|V| \times |V|}$. The objective is to learn a function $f : \mathbf{X}, \mathbf{A} \to \mathbf{H}$ where $\mathbf{H} \in \mathbb{R}^{|V| \times d'}$ is the matrix of node embeddings.

A.2.2 . Graph Neural Network Models

Graph Neural Networks can be broadly categorized based on how they propagate and aggregate information. Some popular models include Graph Convolutional Networks (GCNs) [137], Graph Attention Networks (GATs) [247], and Graph Recurrent Neural Networks (GRNNs). Each of these models has its unique way of processing the graph structure and node features to learn meaningful representa-

tions.

A.2.3 . Message Passing

The message-passing framework is a general approach for designing GNNs. In this framework, each node v_i aggregates messages from its neighbors and updates its representation accordingly. This process can be iterated multiple times to allow information to propagate further across the graph. The general message-passing update can be written as:

$$\mathbf{h}_i^{(t)} = \text{UPDATE}\left(\mathbf{h}_i^{(t-1)}, \text{AGGREGATE}\left(\{\mathbf{h}_j^{(t-1)}, \mathbf{e}_{ij} | j \in \mathcal{N}(i)\}\right)\right)$$

where $\mathbf{h}_i^{(t)}$ is the representation of node v_i at iteration t, $\mathcal{N}(i)$ is the set of neighbors of v_i, and UPDATE and AGGREGATE are functions specific to the GNN model used.

A.2.4 . Graph Convolutions

Graph Convolutional Networks (GCNs) extend the concept of convolutions from grid-like data to graph-structured data [137]. The core idea is to perform a convolution operation on the graph by aggregating features from a node's neighbors. The graph convolution operation for a single layer can be written as:

$$\mathbf{H}^{(l+1)} = \sigma\left(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right)$$

where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix with added self-loops, $\hat{\mathbf{D}}$ is the degree matrix of $\hat{\mathbf{A}}$, $\mathbf{H}^{(l)}$ is the node feature matrix at layer l, $\mathbf{W}^{(l)}$ is the trainable weight matrix, and σ is a non-linear activation function.

A.2.5 . Graph Pooling

Graph pooling methods are used to reduce the size of the graph, making it more manageable for tasks like graph classification. Pooling operations aim to downsample the graph by selecting a subset of nodes and edges or by coarsening the graph. Popular graph pooling methods include GraphSAGE pooling [105], Top-K pooling, and DiffPool [265]. These methods aggregate node features and structure information to create a smaller, more informative graph representation.

In conclusion, Graph Neural Networks provide a robust framework for learning from graph-structured

data by leveraging the relational information between nodes and edges. Their ability to perform message passing, graph convolutions, and pooling operations makes them highly effective for a wide range of applications, from social network analysis to biomedical data processing.

A.2.6 . About Hypergraphs

A Hypergraph is a generalization of the graph structure that extends the interaction between instances to a higher level. To describe this complex relationship where an edge can connect to more than two nodes, we define a hypergraph $\mathcal{G} = (V, E)$ as a hypergraph with M vertices and N hyperedges. The hypergraph can then be generated using an incidence matrix $\mathbf{H} \in \mathbb{R}^{N \times M}$. For each vertex i, the vertex degree is defined as $D_{ii} = \sum_{e \in E} H_{ie}$ and the hyperedge degree will be $B_{ee} = \sum_{i \in V} H_{ie}$.

Hypergraph Convolution

This hypergraph can be associated to a feature matrix $\mathbf{X} \in \mathbb{R}^{N \times F}$ where F is the feature dimension of one node. In the context of our study, this node feature will represent the aggregated Resnet-18 features of one cluster. A step of this convolution is defined in [16] as follows:

$$\mathbf{X}^{(l+1)} = \sigma(\mathbf{D}^{-\frac{1}{2}}\mathbf{H}\mathbf{W}\mathbf{B}^{-1}\mathbf{H}^{T}\mathbf{D}^{-\frac{1}{2}}\mathbf{X}^{(l)}\mathbf{P}) \tag{A.1}$$

where $\mathbf{W}$ is the weight matrix, σ a non-linear transformation and $\mathbf{P}$ is the weight matrix between layer l and l+1.

Hypergraph Attention

To build the attention visualization, we use an attention mechanism for hypergraphs described in [16] as:

$$\alpha_{ij} = \frac{exp(\sigma(sim(x_i\mathbf{P}, x_j\mathbf{P})))}{\sum_{k \in \mathcal{N}_i} exp(\sigma(sim(x_i\mathbf{P}, x_k\mathbf{P})))} \tag{A.2}$$

where the similarity function computes similarity between two vertices as follows:

$$sim(x_i, x_j) = \mathbf{a}^{T}[x_i||x_j] \tag{A.3}$$

where $\mathbf{a}$ is a weight vector and $[.||.]$ denotes concatenation.

A.3 . Generalities on Survival Analysis

Survival analysis is a branch of statistics that deals with the analysis of time-to-event data. The primary objective is to model and predict the time until an event of interest occurs, such as death, relapse, or failure. This type of analysis is widely used in various fields including medicine, biology, engineering, and social sciences. The distinctive feature of survival data is the presence of censoring, which occurs when the event has not been observed for some subjects during the study period. Survival analysis methods account for this censoring to provide accurate estimates and predictions.

A.3.1 . Problem Formulation

In survival analysis, the data typically consists of pairs (T_i, δ_i) for $i = 1, \ldots, n$, where T_i is the observed time for the i-th individual and δ_i is an indicator of whether the event occurred ($\delta_i = 1$) or the data is censored ($\delta_i = 0$). The primary goal is to estimate the survival function $S(t) = P(T > t)$, which gives the probability that the event occurs later than time t, and the hazard function $\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$, which describes the instantaneous rate of occurrence of the event at time t.

Formally, let X_i be a vector of covariates for the i-th individual. The objective is to model the relationship between the survival time T_i and the covariates X_i, often through the hazard function.

A.3.2 . Kaplan-Meier Curves

The Kaplan-Meier estimator is a non-parametric statistic used to estimate the survival function from the observed data. It provides a step function that jumps at each event time. The Kaplan-Meier survival function is given by:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where t_i are the ordered event times, d_i is the number of events at time t_i, and n_i is the number of individuals at risk just before t_i. The Kaplan-Meier curve is a plot of $\hat{S}(t)$ against t, providing a visual representation of the survival experience of the cohort under study [132].

Cox Proportional Hazards Model

The Cox proportional hazards model is a semi-parametric model that assesses the effect of covariates on the hazard function. The Cox model assumes that the hazard function for the i-th individual

can be expressed as:

$$\lambda_i(t \mid X_i) = \lambda_0(t) \exp(X_i^\top \beta)$$

where $\lambda_0(t)$ is the baseline hazard function, X_i is the vector of covariates, and β is the vector of coefficients to be estimated. The Cox model does not assume any specific form for $\lambda_0(t)$, making it flexible in modeling various types of hazard functions. The coefficients β are estimated using partial likelihood, which maximizes the likelihood of the observed ordering of event times [60].

A.3.3 . Evaluation Metrics

Evaluating the performance of survival models involves metrics that can handle censored data. Two commonly used metrics are the concordance index (C-index) and the Brier score.

Concordance Index (C-index)

The C-index is a measure of the model's discriminative power, i.e., its ability to correctly rank the survival times based on predicted risk scores. It is defined as:

$$C = \frac{\sum_{i,j} I(\hat{h}_i > \hat{h}_j) I(T_i < T_j) \delta_i}{\sum_{i,j} I(T_i < T_j) \delta_i}$$

where $\hat{h}_i$ and $\hat{h}_j$ are the predicted risk scores for individuals i and j, I is the indicator function, and δ_i indicates whether T_i is an observed event. A C-index of 0.5 indicates random prediction, while a C-index of 1 indicates perfect prediction [109].

Brier Score

The Brier score measures the accuracy of probabilistic predictions and is adapted for survival analysis to account for censoring. It is defined as:

$$BS(t) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{S}(t \mid X_i) - I(T_i > t) \right)^2$$

where $\hat{S}(t \mid X_i)$ is the predicted survival probability for individual i at time t. The Integrated Brier Score (IBS) is obtained by integrating the Brier score over a range of time points:

$$IBS = \frac{1}{\tau} \int_0^\tau BS(t)\, dt$$

where τ is a pre-specified time horizon [33].

In conclusion, survival analysis provides a comprehensive framework for modeling time-to-event data, accommodating censored observations through various methods. Core concepts like Kaplan-Meier curves, Cox proportional hazards models, and evaluation metrics like the C-index and Brier score are essential for understanding and applying survival analysis in practice.

A.4 . Variational Autoencoders

Variational Autoencoders (VAEs) are a class of generative models that combine principles from deep learning and Bayesian inference. They are designed to model complex distributions and generate new data points similar to those in the training dataset. Unlike traditional autoencoders, which map inputs directly to a latent space and back, VAEs introduce a probabilistic approach to latent variable modeling.

A.4.1 . Key Components of VAEs

- **Encoder Network**: This neural network maps an input data point $\mathbf{x}$ to a distribution over the latent space $q_\phi(\mathbf{z}|\mathbf{x})$. The output is typically a mean and a variance, parameterizing a Gaussian distribution.

- **Latent Space**: The latent variables $\mathbf{z}$ are sampled from the distribution $q_\phi(\mathbf{z}|\mathbf{x})$.

- **Decoder Network**: This neural network maps the latent variable $\mathbf{z}$ back to a distribution over the input space $p_\theta(\mathbf{x}|\mathbf{z})$.

A.4.2 . The Variational Lower Bound (ELBO)

To train VAEs, we aim to maximize the likelihood of the data $p_\theta(\mathbf{x})$. However, directly optimizing this likelihood is intractable due to the integral over the latent variables. Instead, VAEs maximize a variational lower bound on the log likelihood, known as the Evidence Lower Bound (ELBO).

The log likelihood of the data can be decomposed as follows:

$$\log p_\theta(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\right] \tag{A.4}$$

The second term is the Kullback-Leibler (KL) divergence between the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$, which is always non-negative. This gives us:

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \tag{A.5}$$

The ELBO is thus:

$$\mathcal{L}(\theta,\phi;\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})) \tag{A.6}$$

A.4.3 . Deriving the ELBO Loss

To derive the ELBO, we start with the marginal likelihood of the data:

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x},\mathbf{z})\, d\mathbf{z} = \int p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z})\, d\mathbf{z} \tag{A.7}$$

Given the intractability of this integral, we introduce the variational approximation $q_\phi(\mathbf{z}|\mathbf{x})$ and use Jensen's inequality:

$$\log p_\theta(\mathbf{x}) = \log \int q_\phi(\mathbf{z}|\mathbf{x}) \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\, d\mathbf{z} \geq \int q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\, d\mathbf{z} \tag{A.8}$$

This simplifies to the ELBO:

$$\mathcal{L}(\theta,\phi;\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})) \tag{A.9}$$

- **Reconstruction Term**: The first term, $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]$, measures how well the decoder reconstructs the input data from the latent variables.

- **KL Divergence Term**: The second term, $\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}))$, measures the divergence between the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior $p_\theta(\mathbf{z})$.

The ELBO is maximized to train the VAE, ensuring a balance between accurate reconstruction and regularization of the latent space.

By maximizing the ELBO, VAEs learn both to encode data into a meaningful latent space and to generate new data samples that resemble the training data.

A.5 . SHAP Values: Mathematical Framework and Computation

SHAP (SHapley Additive exPlanations) values are a powerful tool for interpreting the output of machine learning models by attributing the contribution of each feature to the model's predictions. Rooted in cooperative game theory, SHAP values are derived from the concept of Shapley values, which were originally designed to distribute payouts fairly among players based on their contribution to the total payout.

A.5.1 . Mathematical Framework

The SHAP value for a feature is computed as the average marginal contribution of that feature across all possible feature combinations. Given a model f and a set of features $X = \{x_1, x_2, \ldots, x_n\}$, the SHAP value ϕ_i for feature x_i is defined as:

$$\phi_i = \sum_{S \subseteq X \setminus \{x_i\}} \frac{|S|!(|X| - |S| - 1)!}{|X|!} \left(f(S \cup \{x_i\}) - f(S)\right)$$

Here:

- S represents a subset of all features excluding x_i.
- $f(S \cup \{x_i\})$ is the model's prediction when feature x_i is included in the subset S.
- $f(S)$ is the model's prediction when feature x_i is excluded from the subset S.
- The term $\frac{|S|!(|X|-|S|-1)!}{|X|!}$ is a weighting factor that accounts for the different permutations of features.

This equation effectively computes the weighted average of the changes in the prediction when the feature x_i is added to every possible subset of features. The SHAP value, therefore, captures the importance of feature x_i by considering all possible interactions with other features.

A.5.2 . Computation of SHAP Values

Computing SHAP values exactly using the above formula can be computationally expensive, especially for models with a large number of features, as it involves evaluating the model on all possible subsets of features. However, several efficient algorithms have been developed to approximate SHAP values, making them feasible for practical use.

Kernel SHAP

This is a model-agnostic approach that approximates SHAP values using a weighted linear regression. It samples subsets of features and fits a weighted linear model to estimate the contributions of each feature.

Tree SHAP

Specifically designed for tree-based models, this algorithm exploits the structure of decision trees to compute SHAP values efficiently. It uses dynamic programming to traverse the tree and calculate the contributions of each feature without needing to evaluate all possible subsets.

Deep SHAP

This method adapts SHAP values for deep learning models by combining ideas from DeepLIFT (Deep Learning Important FeaTures) and Shapley values. It backpropagates contributions through the network layers to estimate feature importance.

A.5.3 . Example Calculation

Consider a simple linear model $f(x) = w_0 + w_1x_1 + w_2x_2$ with two features, x_1 and x_2. To compute the SHAP value for x_1:

1. **Subset** $S = \{\}$: The model prediction without x_1 is $f(\{\}) = w_0$.
2. **Subset** $S = \{x_2\}$: The model prediction with x_2 is $f(\{x_2\}) = w_0 + w_2x_2$.

The marginal contributions for x_1 are:

- When added to $S = \{\}$: $f(\{x_1\}) - f(\{\}) = (w_0 + w_1x_1) - w_0 = w_1x_1$.
- When added to $S = \{x_2\}$: $f(\{x_1, x_2\}) - f(\{x_2\}) = (w_0 + w_1x_1 + w_2x_2) - (w_0 + w_2x_2) = w_1x_1$.

Averaging these contributions, the SHAP value for x_1 is:

$$\phi_1 = \frac{1}{2}w_1x_1 + \frac{1}{2}w_1x_1 = w_1x_1$$

This example demonstrates how SHAP values attribute the model's output to each feature, providing a clear and interpretable measure of feature importance.

A.5.4 . Insights and Applications

SHAP values offer several advantages in model interpretability:

- **Consistency**: They provide consistent and fair feature attributions.
- **Local Interpretability**: They explain individual predictions by decomposing them into contributions from each feature.
- **Global Interpretability**: Aggregating SHAP values across multiple samples provides insights into the overall importance of features in the model.

By leveraging SHAP values, researchers and practitioners can gain a deeper understanding of their models, ensuring transparency and trustworthiness in their predictions.

Datasets

B.1 . The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) is a landmark project initiated by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) aimed at cataloging genetic mutations responsible for cancer. Since its launch in 2006, TCGA has provided an extensive dataset that significantly enhances our understanding of cancer genomics, facilitating the development of improved diagnostics, treatments, and preventive strategies.

TCGA has collected and analyzed tumor samples from over 11,000 patients across more than 33 different cancer types. Each type is represented by a cohort, a group of patient samples studied to unveil the unique genetic and molecular characteristics of that cancer. Major cohorts include Breast Invasive Carcinoma (BRCA), Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), Colon Adenocarcinoma (COAD), Glioblastoma Multiforme (GBM), Ovarian Serous Cystadenocarcinoma (OV), Prostate Adenocarcinoma (PRAD), Skin Cutaneous Melanoma (SKCM), Thyroid Carcinoma (THCA), and Kidney Renal Clear Cell Carcinoma (KIRC). These cohorts collectively provide a detailed landscape of genetic alterations, gene expression patterns, epigenetic modifications, and other molecular features defining each cancer type.

To generate its comprehensive dataset, TCGA utilized a variety of advanced technologies and methodologies. High-throughput sequencing technologies, including whole genome sequencing (WGS) and whole exome sequencing (WES), were employed to identify somatic mutations, copy number variations, and structural rearrangements in cancer genomes. RNA sequencing (RNA-Seq) was used to

analyze gene expression profiles, providing insights into transcriptional activity within cancer cells. DNA methylation profiling techniques, such as Infinium HumanMethylation450 BeadChip and Bisulfite sequencing, assessed DNA methylation patterns, contributing to the understanding of epigenetic changes in cancer. MicroRNA sequencing (miRNA-Seq) was employed to profile microRNA expression, which plays a crucial role in the post-transcriptional regulation of gene expression. Comparative genomic hybridization (CGH) and single nucleotide polymorphism (SNP) arrays were utilized to identify genomic regions with copy number gains or losses, while reverse-phase protein array (RPPA) technology quantified the expression levels of key proteins and phosphoproteins, aiding in the understanding of signaling pathways involved in cancer. Additionally, TCGA integrated clinical and histopathological data with molecular profiles, providing a comprehensive view of each cancer type.

The data generated by TCGA is publicly available and has been instrumental for researchers worldwide. It includes genomic data such as raw sequencing reads, processed mutation calls, and copy number alterations; transcriptomic data from RNA-Seq and miRNA-Seq; epigenomic data with DNA methylation profiles; proteomic data from RPPA; clinical data including detailed patient demographics, treatment protocols, and outcomes; and histopathological data comprising digital images of stained tissue sections and related annotations. Researchers can access TCGA data through various platforms such as the Genomic Data Commons (GDC) Data Portal, the cBioPortal for Cancer Genomics, and the UCSC Cancer Genomics Browser.

B.2 . The MDS Dataset

B.3 . The International Adjuvant Lung Cancer Trial

The International Adjuvant Lung Cancer Trial (IALT) was a pivotal study that evaluated the efficacy of adjuvant chemotherapy in patients with resected non-small cell lung cancer (NSCLC). Conducted between 1995 and 2000, the trial involved 1,867 patients from 33 countries who were randomized to receive either cisplatin-based chemotherapy or no further treatment after surgical resection of their tumors. The results, published in 2004, demonstrated a significant improvement in overall survival for patients who received chemotherapy compared to those who did not, establishing adjuvant chemotherapy as a standard treatment for resected NSCLC [12].

Table B.1: ***Datasets Description*** *Description of the TCGA datasets used in this study for both classification and survival tasks. We show the number of patients available for each modality along with the censoring rate of the cohort.*

		Missing	Mutations	CNV	RNAseq	Methylation	WSI	Overall
n								19188
program, n (%)	TARGET	0						4666 (24.3)
	TCGA							14522 (75.7)
sample_type, n (%)	Additional - New Primary	71						11 (0.1)
	Additional Metastatic							2 (0.0)
	Blood Derived Normal							2 (0.0)
	Bone Marrow Normal							845 (4.4)
	Buccal Cell Normal							5 (0.0)
	FFPE Scrolls							10 (0.1)
	Human Tumor Original Cells							2 (0.0)
	Metastatic							411 (2.1)
	Primary Blood Derived Cancer - Bone Marrow							1054 (5.5)
	Primary Blood Derived Cancer - Peripheral Blood							616 (3.2)
	Primary Tumor							13060 (68.3)
	Recurrent Blood Derived Cancer - Bone Marrow							208 (1.1)
	Recurrent Blood Derived Cancer - Peripheral Blood							14 (0.1)
	Recurrent Tumor							103 (0.5)
	Solid Tissue Normal							2774 (14.5)
project_id, n (%)	TARGET-ALL-P3	234	0	0	0	0	0	341 (1.8)
	TARGET-AML		0	0	0	0	0	2029 (10.7)
	TARGET-CCSK		0	0	0	0	0	19 (0.1)
	TARGET-NBL		0	0	0	0	0	1109 (5.9)
	TARGET-OS		0	0	0	0	0	288 (1.5)
	TARGET-RT		0	0	0	0	0	61 (0.3)
	TARGET-WT		0	0	0	0	0	789 (4.2)
	TCGA-ACC		89	89	80	80	227	97 (0.5)
	TCGA-BLCA		408	408	426	431	457	454 (2.4)
	TCGA-BRCA		1082	1082	1179	881	1133	1283 (6.8)
	TCGA-CESC		284	284	299	299	279	317 (1.7)
	TCGA-CHOL		36	36	45	45	39	71 (0.4)
	TCGA-COAD		442	442	436	335	459	571 (3.0)
	TCGA-DLBC		47	47	46	47	44	52 (0.3)
	TCGA-ESCA		184	184	197	201	158	251 (1.3)
	TCGA-GBM		606	606		151	860	671 (3.5)
	TCGA-HNSC		523	523	568	579	472	612 (3.2)
	TCGA-KICH		65	65	89	65	121	184 (1.0)
	TCGA-KIRC		532	532	587	478	519	985 (5.2)
	TCGA-KIRP		284	284	321	316	300	381 (2.0)
	TCGA-LAML		170	170	164	120		697 (3.7)
	TCGA-LGG		528	528	525	529	844	538 (2.8)
	TCGA-LIHC		372	372	419	424	379	469 (2.5)
	TCGA-LUAD		508	508	541	482	541	877 (4.6)
	TCGA-LUSC		490	490	511	399	512	765 (4.0)
	TCGA-MESO		85	85	85	85	87	88 (0.5)
	TCGA-OV		587	587	494	10	107	758 (4.0)
	TCGA-PAAD		184	184	182	194	209	223 (1.2)
	TCGA-PCPG		169	169	187	187	196	189 (1.0)
	TCGA-PRAD		502	502	551	553	449	623 (3.3)
	TCGA-READ		157	157	156	98	166	192 (1.0)
	TCGA-SARC		261	261	260	266	600	290 (1.5)
	TCGA-SKCM		458	458	439	461	475	477 (2.5)
	TCGA-STAD		408	408	444	382	442	544 (2.9)
	TCGA-TGCT		139	139	139	139	254	156 (0.8)
	TCGA-THCA		511	511	572	570	519	615 (3.2)
	TCGA-THYM		123	123	125	125	181	139 (0.7)
	TCGA-UCEC		538	538	558	465	566	606 (3.2)
	TCGA-UCS		54	54	55	55	91	63 (0.3)
	TCGA-UVM		80	80	80	80	80	80 (0.4)
Age at Diagnosis in Years, mean (SD)		511						46.5 (26.0)
Gender, n (%)	Female	450						9386 (48.9)
	Male							9350 (48.7)
	Unknown							1 (0.0)
	not reported							5 (0.0)

Table B.2: ***MDS Dataset Overview:*** *Overview of the MDS cohort*

Variable	Category	Missing	Overall
n			556
CENTER, n (%)			
	FI	0	29 (5.2)
	KI		510 (91.7)
	MLL		17 (3.1)
PAT_GENDER, n (%)			
	F	3	224 (40.5)
	M		329 (59.5)
R_IPSS_BLAST, mean (SD)		42	1.2 (1.1)
R_IPSS_HB, mean (SD)		74	0.6 (0.7)
WHO_2016, n (%)			
	AML_MRC	0	24 (4.3)
	CMML_0		14 (2.5)
	CMML_1		43 (7.7)
	CMML_2		34 (6.1)
	JMML		1 (0.2)
	MDS_EB_1		55 (9.9)
	MDS_EB_2		65 (11.7)
	MDS_MLD		101 (18.2)
	MDS_MPN		1 (0.2)
	MDS_MPN_RS_T		8 (1.4)
	MDS_MPN_U		10 (1.8)
	MDS_RS_MLD		90 (16.2)
	MDS_RS_SLD		46 (8.3)
	MDS_SLD		11 (2.0)
	MDS_UNS		10 (1.8)
	MDS_del5q		10 (1.8)
	NBM		26 (4.7)
	Z_NOT_MDS		6 (1.1)
	aCML		1 (0.2)
SUR_OS_cens_Event, n (%)			
	0.0	40	268 (51.9)
	1.0		248 (48.1)
SUR_OS_cens, mean (SD)		40	899.1 (863.6)
SUR_H_Blast_P_cens, mean (SD)		250	1619.5 (1188.4)
SUR_H_Blast_P_cens_Event, n (%)			
	0.0	250	287 (93.8)
	1.0		19 (6.2)
SUR_AML_t_cens_Event, n (%)			
	0.0	67	456 (93.3)
	1.0		33 (6.7)
SUR_AML_cens, mean (SD)		67	1548.7 (1214.0)
EPO_duration, mean (SD)		399	639.3 (705.1)
AZA_duration, mean (SD)		390	261.0 (329.1)
SUR_EPO_cens, mean (SD)		399	639.3 (705.1)
SUR_EPO_cens_Event, mean (SD)		0	0.3 (0.5)
SUR_AZA_cens, mean (SD)		390	261.0 (329.1)
SUR_AZA_cens_Event, n (%)			
	0	0	390 (70.1)
	1		166 (29.9)

Table B.3: ***IALT Overview*** *Overview of the IALT cohort with details about the different clinical and molecular features involved.*

Variable	Category	Missing	Overall
n			544
Age, n (%)			
	55-64	0	241 (44.3)
	< 55		160 (29.4)
	>=65		143 (26.3)
Sex, n (%)			
	Female	0	114 (21.0)
	Male		430 (79.0)
PS, n (%)			
	0	0	276 (50.7)
	1-2		268 (49.3)
Surgery, n (%)			
	Lobectomy/Other	0	324 (59.6)
	Pneumonectomy		220 (40.4)
T, n (%)			
	T1	0	85 (15.6)
	T2		340 (62.5)
	T3/T4		119 (21.9)
N, n (%)			
	No	0	245 (45.0)
	N1		166 (30.5)
	N2		133 (24.4)
Arm, n (%)			
	Chemotherapy	0	277 (50.9)
	Control		267 (49.1)
Histology, n (%)			
	Adenocarcinoma	0	186 (34.2)
	Other		61 (11.2)
	Squamous Cell Carcinoma		297 (54.6)
stime, mean (SD)		0	1608.7 (1052.5)
status, n (%)			
	Alive	0	208 (38.2)
	Dead		336 (61.8)
dfstime, mean (SD)		0	1016.8 (706.7)
dfs, n (%)			
	Any event (relapse or death)	0	310 (57.0)
	No event		234 (43.0)
dcause, n (%)			
	Alive	0	271 (49.8)
	Chemotherapy toxicity		2 (0.4)
	Other		37 (6.8)
	Progression		213 (39.2)
	Unknown		21 (3.9)
tnfail, n (%)			
	No	0	441 (81.1)
	Yes		103 (18.9)
mfail, n (%)			
	No	0	368 (67.6)
	Yes		176 (32.4)
MAPD, mean (SD)		0	0.3 (0.1)
ndSNPQC, mean (SD)		0	23.5 (8.8)

Supplementary Materials

C.1 . Multimodal Representation Learning for High-Dimensional Data

C.2 . Multi-Omics Integration for Precision Medicine

C.2.1 . CustOmics: A versatile deep-learning based strategy for multi-omics integration

Table C.1: *Number of trainable parameters for each model used for the downstream tasks*

Method	Number of Parameters
Early Int. VAE	$4,754,963$
Joint Int. VAE	$4,997,702$
Late Int. VAE	$5,045,478$
Mix Int. AE	$4,457,236$
CustOmics	$5,112,119$

C.2.2 . An Application of Multi-Omics Integration to Myelodysplastic Syndromes

Model	Accuracy	Macro F1-Score	AUC
CustOmics	0.9223 $\pm$ 0.0157	0.9221 $\pm$ 0.0134	0.9245 $\pm$ 0.0098
OmiEmbed	0.8901 $\pm$ 0.0143	0.8905 $\pm$ 0.0118	0.8912 $\pm$ 0.0105
MFA	0.8857 $\pm$ 0.0172	0.8720 $\pm$ 0.0141	0.8808 $\pm$ 0.0112
Logistic Regression	0.8012 $\pm$ 0.0185	0.7987 $\pm$ 0.0159	0.8123 $\pm$ 0.0121

Table C.2: *Comparison of multi-omics integration models for MDS vs CMML classification*

C.3 . Analysis of Histopathology Slides

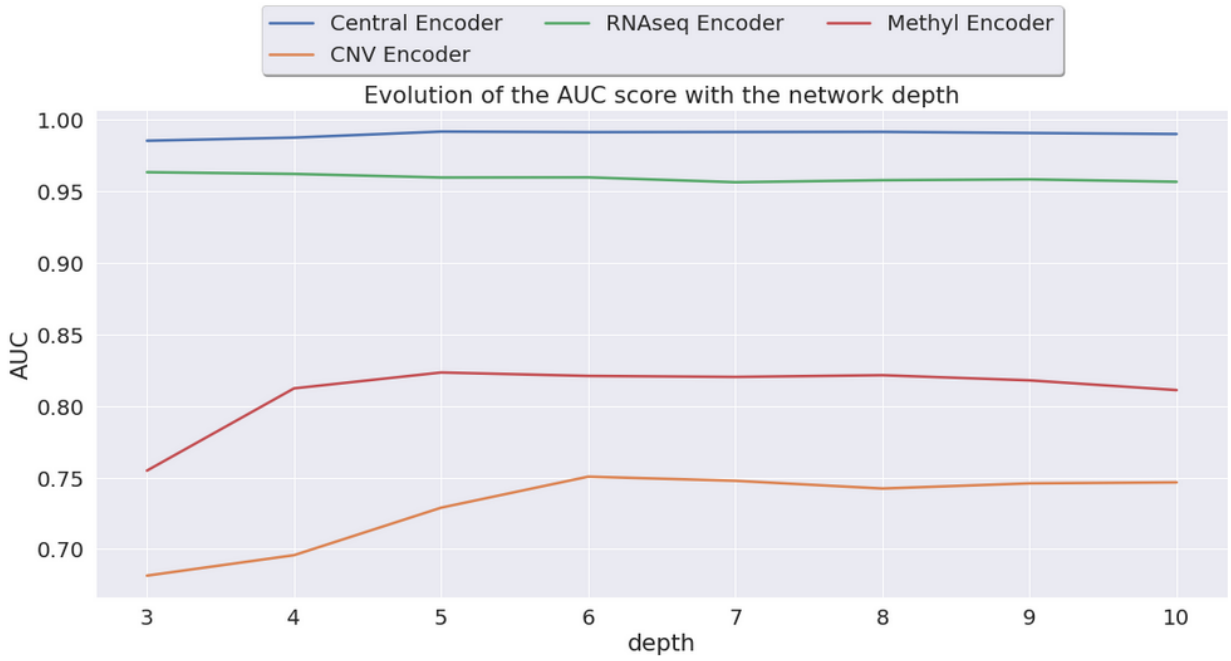


Figure C.1: ***Evolution of the AUC score with the network's depth:*** *We first assess the evolution of the performance on the tumor classification task for each source using the intermediate autoencoders, then we evaluate the effect of the depth on the central encoder using the best results for the intermediate autoencoders for each source. We see that RNAseq data does not need as many layers as CNV and methylation data, suggesting that its convergence may be simpler as it holds most of the signal for tumor-type prediction.*

C.3.1 . Hyper-AdaC: Adaptive clustering-based hypergraph representation of whole slide images for survival analysis

Patch Clustering

We compute the average number of elements remaining after the hierarchical clustering step for each dataset separately, the results along with the ratio between initial and filtered patches are represented in Table C.3. We observe that, in general, approximately 14% of the WSI is used (see C.3), and as shown in Figure 4.6, those elements are well spread across the WSI. However, we can see that both BLCA and GBMLGG datasets behave differently from the others. For BLCA, the ratio of remaining elements over the total number of patches is higher than all the other datasets, whereas for GBMLGG it is the opposite. Our method does not perform well for those particular test cases.

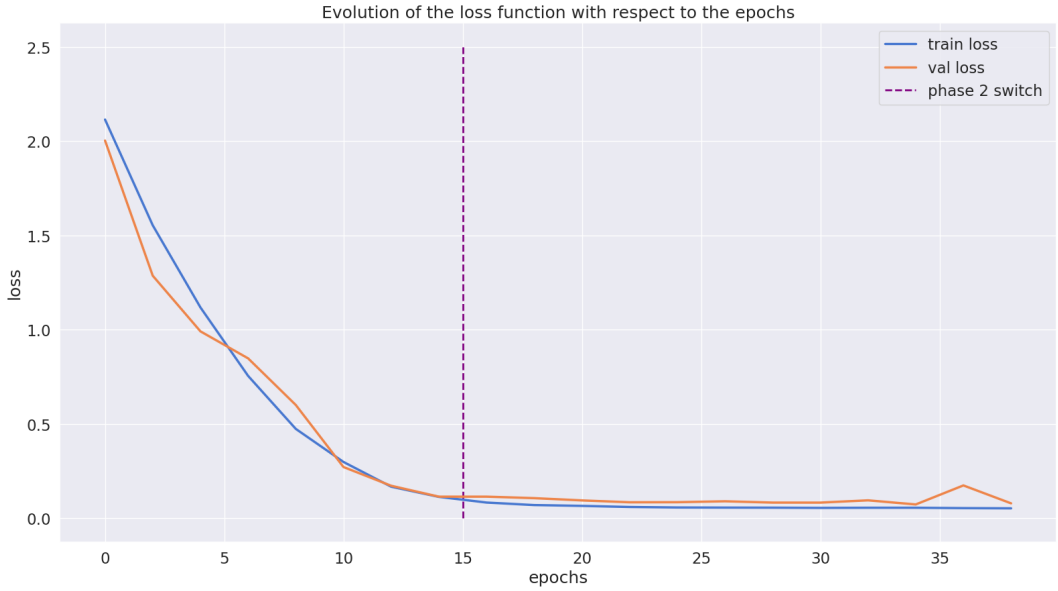


Figure C.2: ***Evolution of the loss function:*** *We display the evolution of both training and validation losses before and after the phase switch for the tumor classification task.*

Table C.3: *Average number of nodes after the hierarchical clustering step for each dataset. Due to our selection criteria, the GBMLGG dataset had a significantly lower number of nodes (as the ratio is also lower, it may indicate higher homogeneity among tissues), which may explain the lower performance with respect to the other cancer types. In this study, we selected the same hyperparameters for all the cancer types to prove the generalizability of our method, outperforming the other state-of-the-art methods. Some specific hyperparameters tuning for the GBMLGG and BLCA may resolve this issue.*

Cancer Type	**# of patches**	**# of nodes**	$\frac{\text{\# of nodes}}{\text{\# of patches}}$
Bladder Urothelial Carcinoma (BLCA)	58586	9187	0.16
Breast Invasive Carcinoma (BRCA)	38107	5304	0.14
Glioblastoma & Lower Grade Glioma (GBMLGG)	15855	961	0.06
Lung Adenocarcinoma (LUAD)	43445	6003	0.14
Ulterine Corpus Endometrial Carcinoma (UCEC)	56162	7748	0.14

Ablation Study

We perform an ablation study on the different graph hyperparameters to justify our construction choices. In Figure C.5, we can see the effect of the similarity threshold δ_h on the survival performances. The stricter the constraint, the better the performance, indicating that larger graphs fail at learning generalizable properties. This idea is also supported by the standard deviation across the 5-folds that decreases, suggesting that the model is less robust with larger graphs. A similarity threshold of 80%

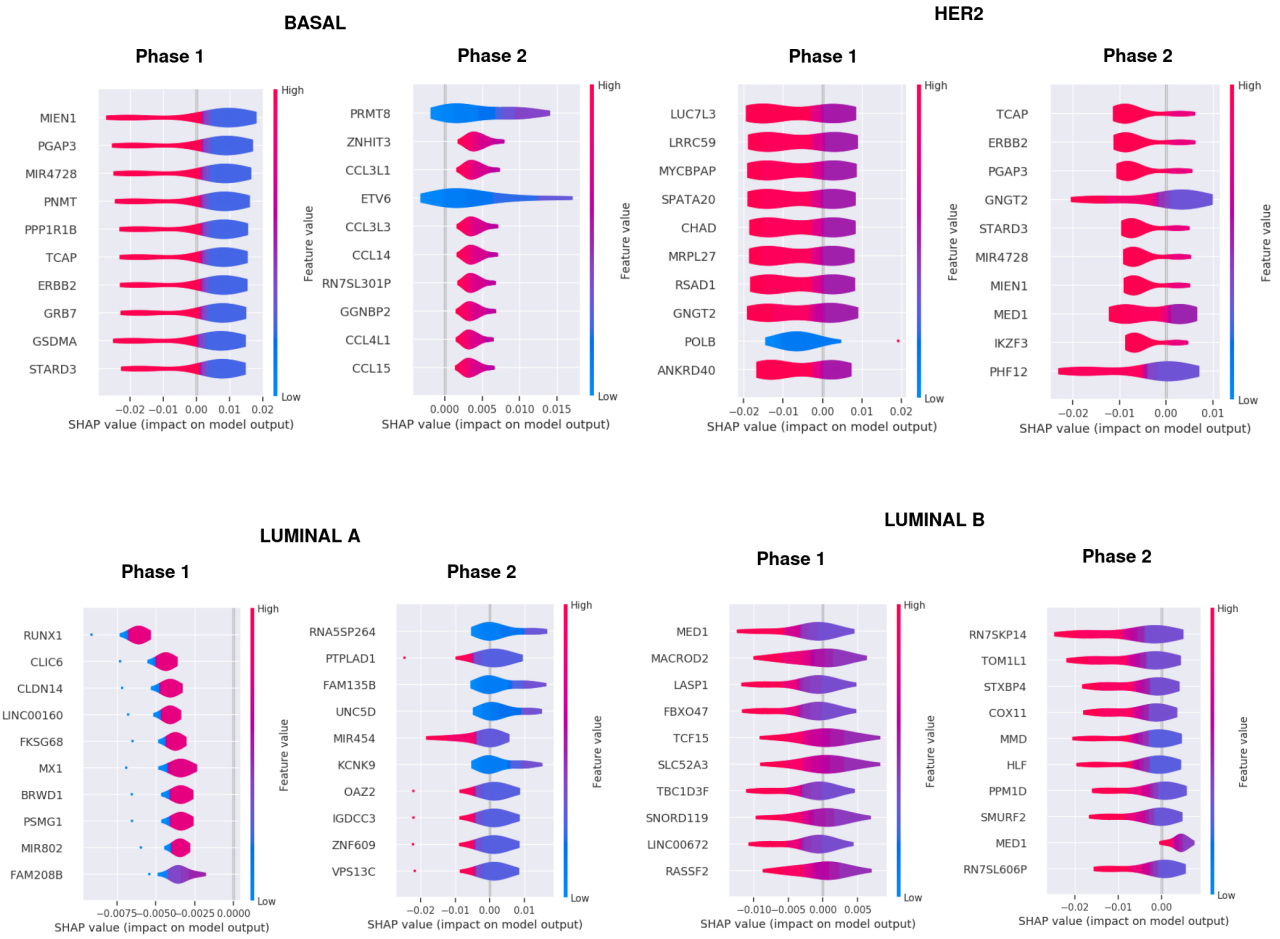


Figure C.3: *PAM50 gene importance: Computed SHAP values on the CNV data of the most relevant genes responsible for discriminating between subtypes against the others using CustOmics for both integration phases.*

achieves the peak performance; past that point, the performances start to decrease again because we tend to oversimplify the WSI and start neglecting information.

Figure C.6 highlights the relationship between morphological features and geographical properties with respect to the survival prediction performance. We see that, in general, focusing on morphological properties is more beneficial to the performances than spatial properties as they hold more information about the structure of the tissue (including, to a certain extent, spatial information because similar patches tend to be close). However, focusing too much on morphological features can hinder the accuracy of our survival predictions, as sometimes the homogeneity of specific tissues can make the filtering biased and overlook chunks of WSIs that may hold vital information.

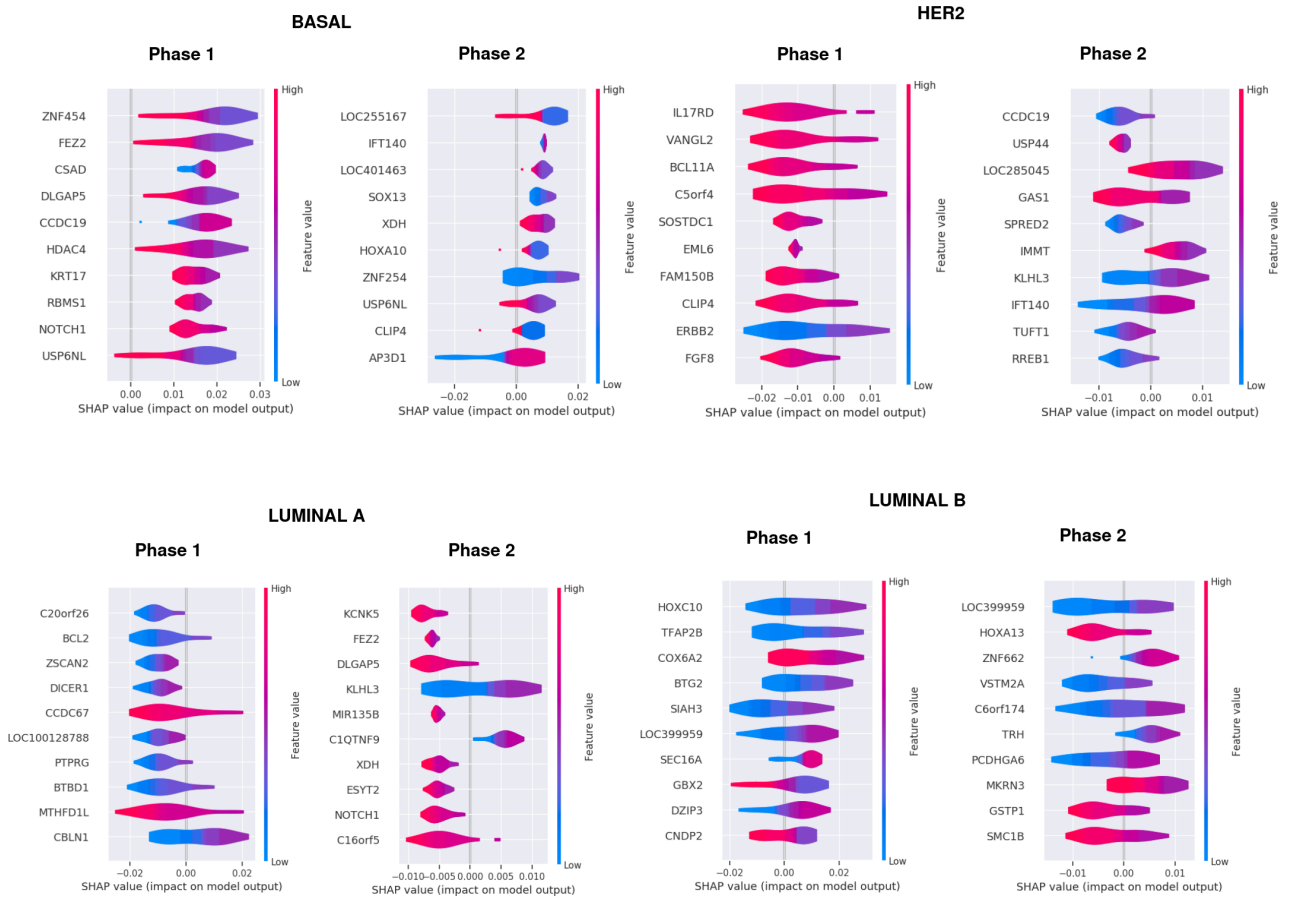


Figure C.4: *PAM50 gene importance: Computed SHAP values on the methylation data of the most relevant genes responsible for discriminating between subtypes using CustOmics for both integration phases.*

C.4 . Multimodal Integration of Multi-Omics Data & Histopathology Slides

Table C.4: ***Modality Combinations*** *Performance comparison between multiple combination of modalities for CustOmics for the Pancancer classification task. The evaluation is done using the Area Under ROC-curve (AUC).*

Omics Combinations	SNN	CustOmics
CNV	74.3 ± 3.0	75.1 ± 2.7
RNAseq	94.0 ± 2.6	96.0 ± 1.4
Methyl	81.1 ± 1.7	82.3 ± 1.3
CNV + RNAseq	94.3 ± 2.9	96.9 ± 0.8
CNV + Methyl	81.4 ± 1.8	85.7 ± 2.1
RNAseq + methyl	93.2 ± 1.0	97.3 ± 0.7
CNV + RNAseq + Methyl	94.1 ± 2.7	98.9 ± 1.4

Table C.5: ***Ablation Study*** *Performance comparison between CustOmics and the state of the art for classification tasks by replacing different instances of the model:* ***a. Multi-Omics Ablation:*** *CustOmics A1 replaces the multi-omics VAE with an SNN.* ***b. Hypergraph Ablation:*** *CustOmics A2 replaces the hypergraph encoder with a visual transformer and CustOmics A3 replaces the hypergraph represnetation with a regular graph embedding.* ***c. Downstream Network Ablation:*** *CustOmics A4 replaces the hierarchical mixture-of-experts approach with a regular mixture of experts network, CustOmics A5 replaces it with a transformer classifier.*

Methods	PANCAN	BRCA	COAD	STAD
SNN (Multi-Omics)	94.1 ± 2.7	92.0 ± 3.3	79.2 ± 3.3	84.6 ± 3.3
CustOmics (Multi-Omics)	$\mathbf{98.9 \pm 1.4}$	$\mathbf{98.3 \pm 1.0}$	$\mathbf{88.1 \pm 1.9}$	$\mathbf{98.4 \pm 1.2}$
DeepSets (WSI Only)	84.7 ± 3.3	68.6 ± 4.1	55.2 ± 4.3	58.9 ± 4.6
AttnMIL (WSI Only)	88.4 ± 2.0	71.2 ± 4.9	56.2 ± 4.4	61.4 ± 4.4
DeepAttnMISL (WSI Only)	89.8 ± 2.5	71.1 ± 3.3	55.7 ± 4.0	62.1 ± 4.5
MCAT (WSI Only)	90.4 ± 1.8	72.3 ± 3.3	61.7 ± 3.3	69.4 ± 3.5
SurvPATH (WSI Only)	89.7 ± 2.4	71.4 ± 3.6	60.2 ± 4.1	69.1 ± 3.2
CustOmics A1 (WSI Only)	85.1 ± 3.8	70.5 ± 5.5	57.7 ± 4.1	59.4 ± 4.4
CustOmics A2 (WSI Only)	90.4 ± 1.8	72.3 ± 3.3	61.7 ± 3.3	69.4 ± 3.5
CustOmics A3 (WSI Only)	88.7 ± 1.2	69.4 ± 3.6	55.1 ± 3.5	61.9 ± 5.0
CustOmics A4 (WSI Only)	89.4 ± 2.1	71.0 ± 3.3	58.8 ± 3.3	67.2 ± 3.7
CustOmics A5 (WSI Only)	87.1 ± 1.9	70.4 ± 3.6	55.2 ± 3.9	66.4 ± 4.1
DeepSets (WSI + Multi-Omics)	96.7 ± 1.5	84.9 ± 2.0	58.6 ± 2.2	67.1 ± 2.7
AttnMIL (WSI + Multi-Omics)	97.1 ± 1.2	86.6 ± 2.1	60.0 ± 2.5	69.4 ± 2.7
DeepAttnMISL (WSI + Multi-Omics)	97.8 ± 1.1	88.3 ± 2.7	65.4 ± 2.7	66.6 ± 2.2
MCAT (WSI + Multi-Omics)	98.9 ± 1.1	95.4 ± 2.0	93.3 ± 1.2	88.7 ± 2.3
SurvPATH (WSI + Multi-Omics)	98.1 ± 1.7	94.8 ± 2.2	93.5 ± 1.1	87.9 ± 2.1
CustOmics A1 (WSI + Multi-Omics)	96.1 ± 1.7	85.2 ± 2.4	59.5 ± 1.9	67.7 ± 2.9
CustOmics A2 (WSI + Multi-Omics)	99.0 ± 1.3	98.4 ± 2.2	94.1 ± 1.2	95.3 ± 2.4
CustOmics A3 (WSI + Multi-Omics)	99.9 ± 1.1	98.2 ± 2.5	93.9 ± 1.9	94.1 ± 2.7
CustOmics A4 (WSI + Multi-Omics)	99.4 ± 0.8	98.0 ± 1.2	94.2 ± 1.1	96.1 ± 2.2
CustOmics A5 (WSI + Multi-Omics)	98.5 ± 1.9	97.3 ± 3.4	93.8 ± 2.0	94.2 ± 2.9

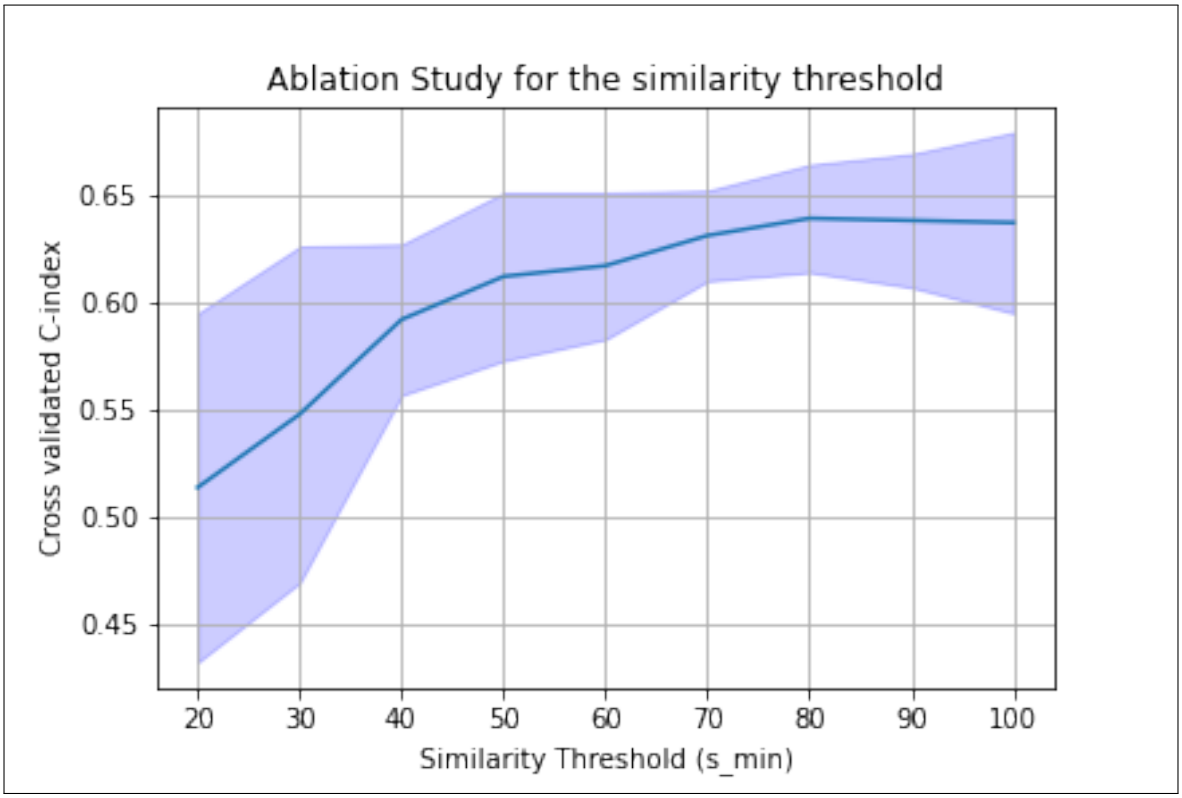


Figure C.5: *Ablation Study for the similarity threshold δ used in the hierarchical clustering step. We evaluate for each hyperparameter the 5-fold cross-validated C-index on the overall 5 TCGA datasets used in this study.*

Table C.6: **Multiple Gene Sets Study Classification:** *CustOmics' classification perfomances across all tasks for multiple gene sets.*

Methods	PANCAN	BRCA	COAD	STAD
Hallmarks	99.5 ± 0.9	98.7 ± 1.1	94.7 ± 1.0	96.3 ± 2.4
Oncologic Signatures	99.5 ± 1.1	98.1 ± 2.8	95.1 ± 1.6	94.8 ± 2.0
Reactome	99.5 ± 1.1	98.1 ± 2.8	95.1 ± 1.6	94.8 ± 2.0

Table C.7: **Multiple Gene Sets Study Survival:** *CustOmics' survival perfomances across all tasks for multiple gene sets.*

Methods	BLCA	BRCA	COAD	GBMLGG	KIRC	LUAD	STAD	UCEC
Hallmarks	67.2 ± 2.5	65.2 ± 3.6	64.5 ± 2.2	84.2 ± 2.3	68.2 ± 2.1	64.9 ± 3.7	58.0 ± 1.5	68.0 ± 2.2
Oncologic Signatures	67.6 ± 2.3	63.2 ± 3.0	63.5 ± 2.6	85.8 ± 4.0	66.2 ± 4.7	62.9 ± 4.2	60.0 ± 4.2	69.4 ± 1.7
Reactome	67.7 ± 4.2	67.2 ± 4.6	62.5 ± 1.9	82.7 ± 2.7	69.4 ± 1.0	66.8 ± 3.0	56.0 ± 1.5	66.1 ± 2.4

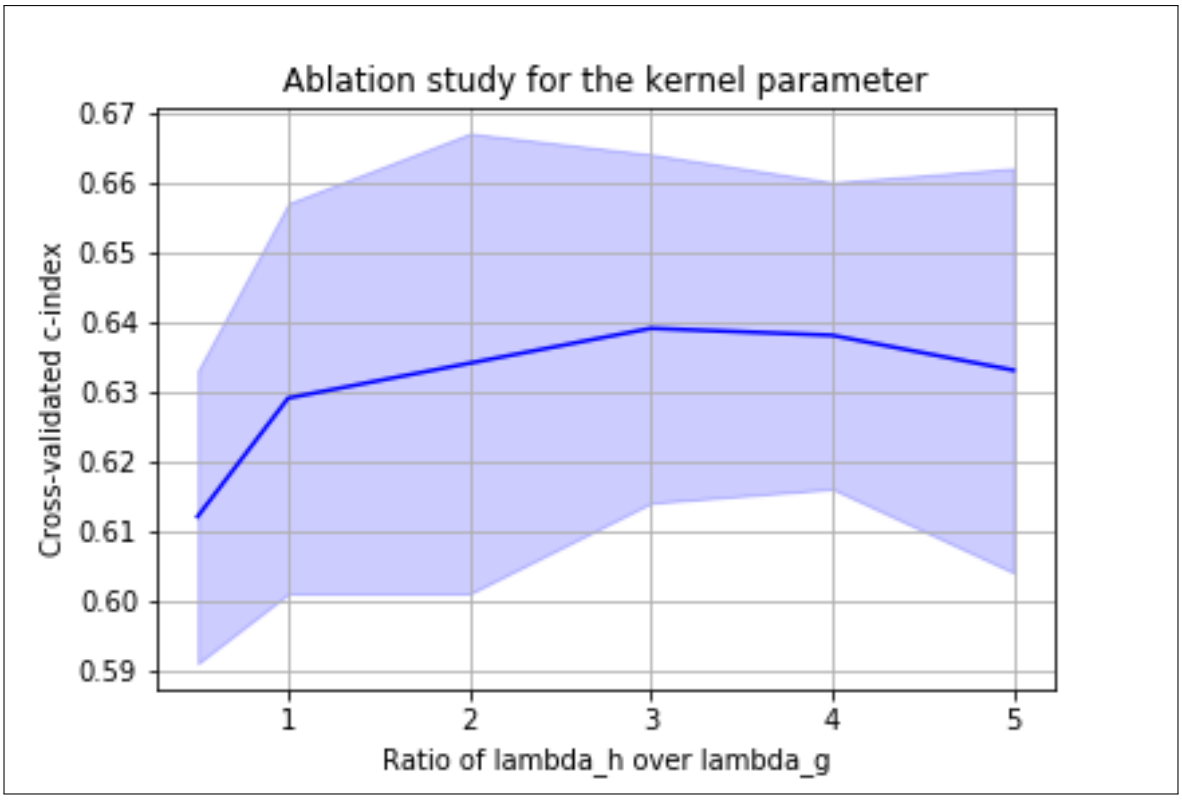


Figure C.6: *Ablation study for $\frac{\lambda_h}{\lambda_g}$ used in the hierarchical clustering step. We evaluate for each hyperparameter the 5-fold cross-validated C-index on the overall 5 TCGA datasets.*