Multimodal Learning with Graphs for Multiomics Cancer Classification

Sina Tabakhi

Supervisor: Prof. Haiping Lu

School of Computer Science

University of Sheffield

This thesis is submitted for the degree of

Doctor of Philosophy

April 2026

Declaration

All sentences or passages quoted in this thesis from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this thesis have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in the degree examination as a whole.

Name: Sina Tabakhi

Signature: Sina Tabakhi

Date: 8 April 2026

Acknowledgements

Pursuing and hopefully completing a PhD has given me the opportunity to grow both intellectually and personally. This thesis is the result of several years of dedicated work, shaped by the contributions of many individuals, whether directly or indirectly involved. Without their guidance, feedback, and support, it would not have been possible.

First and foremost, I would like to express my sincere thanks to my supervisor, Prof. Haiping Lu, for his constant support and dedication throughout my PhD. His regular advice and encouragement have helped me mature in all aspects of research, from formulating problems to presenting and communicating results. I would also like to give special thanks to my collaborators, Dr. Charlotte Vandermeulen, Dr. Ian Sudbery, Dr. Edward C. Harding, and Dr. Florian T. Merkle, whose input has been valuable in shaping my research. I am also grateful to the University of Sheffield and the School of Computer Science for providing the research scholarship and financial support, which made this thesis possible.

Special thanks go to the members of the Machine Learning research group with whom I have shared an office over the years: Dr. Peizhen Bai, Dr. Lawrence Schöbs, Dr. Shuo Zhou, Dr. Xianyuan Liu, Pawel Pukowski, Mohammod N.I. Suvon, Dr. Prasun C. Tripathi, Rea Nkhumise, Alan Thomas, Jiayang Zhang, Lalu M. Riza Rizky, Wenrui Fan, Haolin Wang, and so many more. You are an amazing group of people, and I am grateful for the good times we shared. Your enthusiasm, collaborative spirit, and commitment to advancing knowledge have been a constant source of inspiration.

Most importantly, I am deeply grateful to my parents and sister for always being there for me and for their unwavering support. Their encouragement and understanding throughout this journey have been my anchor. This achievement is dedicated to them with gratitude.

Abstract

With advances in high-throughput technologies, multiple high-dimensional molecular modalities, known as multiomics, have become increasingly available, offering complementary insights into cancer biology. Graph-based multimodal learning has shown remarkable potential in integrating these modalities to unravel cancer complexity, enhance biological predictions, and facilitate biomarker discovery. Despite their promise, these models face three challenges. First, they struggle with small patient cohorts and high-dimensional features, often applying independent feature selection without capturing relationships across omics modalities. Second, conventional graph-based models rely on homogeneous graphs that cannot represent multiple node and edge types. Third, handling missing modalities remains another open problem, as the number of missing patterns increases exponentially with the number of modalities.

This thesis proposes four graph-based multimodal learning models designed to improve accuracy, interpretability, and biomarker discovery in cancer diagnosis. The first model develops a multimodal feature selection method based on a multi-agent system that captures both intra- and inter-omics interactions. Building on this, the second model introduces the automatic construction of heterogeneous graphs from multiomics data to learn holistic, omics-specific representations. To handle datasets with missing modalities, the third model presents a direct prediction approach for partial modalities by introducing a patient-modality multi-head attention mechanism, whose complexity increases linearly with the number of modalities while adapting to missing-pattern variability. Finally, the fourth model extends the proposed multimodal feature selection method to handle missing modalities and is applied to a use case investigating the effects of diet-induced obesity and metformin treatment on

molecular changes in mice. Comprehensive experiments show the superior performance of the proposed methods on real-world cancer datasets and their effectiveness in identifying biologically meaningful biomarkers.

Table of Contents

List of Algorithms

List of Figures

List of Tables

Symbols and Notations

General

$[a,b]$	Closed real interval including a and b
$[a,b)$	Half-open real interval including a but excluding b
$\{x,y,z\}$	Unordered set
$\mathbb{R}$	Space of real numbers
x	Lowercase italic letter denotes a scalar
$\mathbf{x}$	Lowercase bold letter denotes a vector
$\mathbf{X}$	Uppercase bold letter denotes a matrix
X	Uppercase bold italic letter denotes a tensor
$\mathcal{X}$	Uppercase calligraphic letter denotes a set
x_i	i-th element of vector $\mathbf{x}$
x_{ij}	Element at i-th row and j-th column of matrix $\mathbf{X}$
$\mathbf{x}_i$	i-th row of matrix $\mathbf{X}$

Graph Theory

$\mathbf{A}$	Adjacency matrix
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	Homogeneous graph with node set $\mathcal{V}$ and edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$
$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \phi, \psi)$	Heterogeneous graph with node and edge type maps ϕ and ψ
$\mathcal{N}(u)$	Neighbors of node $u \in \mathcal{V}$
$\lvert\mathcal{N}(u)\rvert$	Degree of node $u \in \mathcal{V}$

Multimodal Learning

$\mathcal{D} = \{\mathbf{X}^1, \ldots, \mathbf{X}^M\}$	Multimodal dataset with M modalities
$\mathcal{D}_{\text{tr}}^i$	Training set for modality i
$\mathbf{M} \in \{0,1\}^{N \times M}$	Binary modality mask matrix for missing modalities
$\mathbf{X}^i \in \mathbb{R}^{N \times d_i}$	Modality i with N patients and d_i features
$\mathbf{Y} \in \{0,1\}^{N \times C}$	One-hot encoded label matrix for N patients and C classes
$\hat{\mathbf{Y}}$	Model-predicted label matrix

Functions and Operations

$(\cdot)^\top$	Transpose operator
$\odot$	Element-wise (Hadamard) product
$\|\cdot\|^2$	Squared Euclidean distance
$\exp(\cdot)$	Exponential function
f_θ	Function f parameterized by θ
$\mathrm{KL}(\mathbf{P} \,\|\, \mathbf{Q})$	Kullback–Leibler divergence between distributions $\mathbf{P}$ and $\mathbf{Q}$
$\log(\cdot)$	Natural logarithm
$\mathcal{L}$	Loss function
$\prod$	Product over a sequence of terms
$\sigma(\cdot)$	Nonlinear activation function
$\sum$	Summation over a sequence of terms

Abbreviations

ANN	Artificial Neural Network
AUPRC	Area Under the Precision-Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast Invasive Carcinoma
CNV	Copy Number Variation
DB	Davies-Bouldin Index
DNA	Deoxyribonucleic Acid
DT	Decision Tree
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GIN	Graph Isomorphism Network
GNN	Graph Neural Network
GO	Gene Ontology
GraphSAGE	Graph Sample and Aggregation
HFD	High-Fat Diet
KL	Kullback–Leibler Divergence
KNN	k-Nearest Neighbors
LGG	Lower Grade Glioma
LR	Logistic Regression
Macro F1	Macro-averaged F1 Score
MAS	Multi-Agent System

MCC	Matthews Correlation Coefficient
Micro F1	Micro-averaged F1 Score
MLP	Multilayer Perceptron
NPV	Negative Predictive Value
OV	Ovarian Serous Cystadenocarcinoma
PPV	Positive Predictive Value
RCC	Renal Cell Carcinoma
RF	Random Forest
RGAT	Relational Graph Attention Network
RGCN	Relational Graph Convolutional Network
ROC	Receiver Operating Characteristic
RNA	Ribonucleic Acid
SS	Silhouette Score
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
UMAP	Uniform Manifold Approximation and Projection
VCDN	View Correlation Discovery Network
Weighted F1	Weighted-averaged F1 Score
XGBoost	Gradient Tree Boosting

Chapter 1

Introduction

1.1 Motivation

Information about phenomena in the natural world often comes from multiple modalities. Each modality originates from a different source and represents distinct statistical properties, while multimodal data include related information from various data modalities. Multimodal learning offers new opportunities for developing machine learning models that incorporate complementary information extracted across modalities to solve complex problems (Liu et al., 2025; Baltrušaitis et al., 2018; Xu et al., 2023b). It has shown promise in a variety of domains (Goyal et al., 2023; He et al., 2022; Liu et al., 2024; Rappaport et al., 2020; Leng et al., 2024), with healthcare data analytics being one of the most impactful areas (Steyaert et al., 2023; Xiang et al., 2025; Acosta et al., 2022; Krones et al., 2025).

Multimodal health data analytics leverages multiple biomedical datasets, which have been exponentially produced through technological advances in imaging, sequencing, and molecular profiling (Reel et al., 2021; Kang et al., 2022; Krones et al., 2025), to address real-world healthcare problems. Fine-grained biological omics data from different high-throughput platforms enable scientists to investigate complex life-threatening diseases such as cancer, where development is driven by alterations across multiple molecular layers. In oncology, patient profiling using multiple omics modalities (i.e., genomics, epigenomics, transcriptomics, proteomics, metabolomics, and others) has become increasingly common,

with each modality providing unique value in understanding tumor biology (Ma et al., 2025; Steyaert et al., 2023; Tong et al., 2020; Picard et al., 2021; Mamoshina et al., 2018; Dias-Audibert et al., 2020; Arslan et al., 2021). Collectively, integrating complementary information from the interactions between these omics data (referred to as multiomics analysis) offers a more comprehensive understanding of cancer mechanisms and supports clinical tasks such as classification, disease subtyping, prognosis, and biomarker identification (Acosta et al., 2022; Ding et al., 2022; Krassowski et al., 2020).

With the outstanding efforts of researchers and the vast investments of multiple institutes, notable projects have been completed and made publicly available to the research community. The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) and the International Cancer Genome Consortium (ICGC) (Zhang et al., 2011) have built large-scale multiomics datasets to advance precision medicine. Building on these resources, research initiatives developing deep learning methods for multiomics integration (Cantini et al., 2021; Acosta et al., 2022; Schulte-Sasse et al., 2021) leverage the unique and complementary characteristics of each modality to construct multimodal models. These models are more accurate and interpretable than unimodal models, thereby enhancing biological predictions and facilitating biomarker discovery (Karczewski and Snyder, 2018; Li et al., 2022a; Schulte-Sasse et al., 2021).

More recently, deep learning on graphs (or networks), referred to as graph neural networks (GNNs) (Hamilton et al., 2018), has been increasingly utilized for multiomics integration (Li et al., 2022a; Forster et al., 2022; Wang et al., 2021b; Wu et al., 2024a; Ma et al., 2025). Graphs provide a natural way to represent complex biological systems, where nodes correspond to entities such as genes, proteins, or patients, and edges capture their interactions or relationships (Hamilton, 2020). Modeling each omics modality as a graph facilitates the extraction of structural information within and across datasets, enabling GNNs to effectively capture both intra- and inter-omics dependencies (Li et al., 2022a; Ektefaie et al., 2023). This graph-based representation not only provides a deeper understanding of biological mechanisms but also offers more accurate models with improved decision-making power (Zitnik et al., 2024).

Despite the promise and rich potential of graph-based learning, its application to multiomics analysis faces challenges arising from the complex nature of multiomics data. One major difficulty is the high dimensionality of each omics modality, which makes it challenging to learn meaningful representations (Steyaert et al., 2023; Tabakhi et al., 2023; Cantini et al., 2021). Early attempts to mitigate this issue focused on reducing dimensionality independently for each modality (Picard et al., 2021; Wang et al., 2021b; Wu et al., 2024a); however, such methods fail to account for relationships across modalities. Furthermore, representation learning on multimodal graph-structured data has largely been limited to homogeneous graphs with a single type of node and edge (Zheng et al., 2024; Wang et al., 2021b; Forster et al., 2022; Schulte-Sasse et al., 2021). Only a few GNN methods extend to heterogeneous graphs by modeling multiple types of nodes and edges (Zitnik et al., 2024; Ruiz et al., 2023; Xu et al., 2023a), and these approaches often rely on pre-existing knowledge graphs that may not always be available (Chandak et al., 2023). Another challenge is the presence of missing modalities, where all data from some modalities are missing for some patients (Mitra et al., 2023). While several methods have been proposed to address missing modalities (Hayat et al., 2022; Yao et al., 2024; Tsai et al., 2019; Ma et al., 2021; Reza et al., 2024), they are typically unable to accommodate diverse missing-modality patterns, or multimodal missingness scalability remains a challenge. These limitations hinder the development of powerful graph-based multiomics methods for downstream tasks in healthcare. This thesis aims to address these issues by designing graph-based multiomics integration methods that enable effective cancer classification and subtyping. From a methodological perspective, this involves moving beyond single-modality feature selection toward multimodal feature selection, developing approaches for multimodal heterogeneous graph learning, and designing flexible and efficient models for multimodal learning in the presence of missing data. Figure 1.1 shows the workflow and challenges of multimodal learning with graphs for multiomics integration.

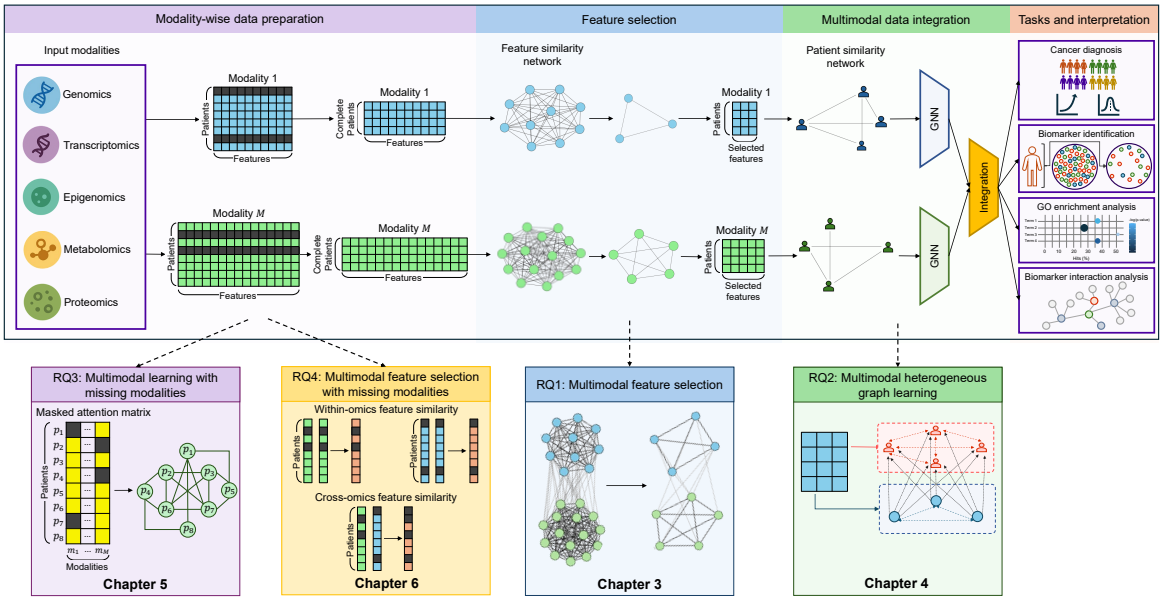

Fig. 1.1 **Workflow of multimodal learning with graphs for multiomics integration and the four research questions addressed in this thesis. Top:** High-dimensional input modalities, which may include missing modalities, first undergo patient filtering to exclude incomplete data, followed by graph-based feature selection to reduce complexity. The selected features are then integrated across modalities through graph learning, enabling downstream tasks and biological interpretation. **Bottom:** The four research questions (RQ1–RQ4) correspond to the methodological developments discussed in Chapters 3–6, illustrating how each contributes to different stages of the overall workflow.

1.2 Research Questions

In this thesis, we investigate the challenges of multimodal learning with graphs for multiomics cancer classification and propose integrative models to address current limitations and improve performance. To this end, we aim to answer the following research questions, and a summary of them is presented in Figure 1.1.

Research Question 1: *How can we develop a multimodal feature selection method that reduces omics feature dimensionality while accounting for both intra- and inter-omics relationships?*

Multiomics data are inherently high-dimensional, with each omics modality containing a large number of features relative to the small patient cohorts (Cantini et al., 2021; Kang

et al., 2022). As a result, feature selection is a common preprocessing step. Feature selection can improve the ability of GNNs to learn meaningful representations; however, most existing approaches apply feature selection independently to each omics modality. This overlooks potential relationships and complementary information across modalities, which could further enhance model performance (El-Manzalawy et al., 2018). A few studies have attempted multimodal feature selection across all omics, but these methods are often limited to single-iteration greedy strategies, leading to reduced predictive capability. This naturally raises the question of how to develop more effective feature selection methods that account for both intra- and inter-omics relationships in multiomics data.

Research Question 2: *How can we construct heterogeneous graphs to learn holistic graph representations that capture diverse structures in multiomics data?*

Most current GNN research in multiomics is limited to homogeneous graphs, where the learning process is built on patient or feature similarity networks with a single type of node and edge (Li et al., 2022a). Such representations lose crucial structural information, overlook the diverse nature of multiomics data, and restrict the model's ability to capture complex biological interactions. Heterogeneous graphs provide a promising alternative by modeling multiple types of nodes and edges, thereby better reflecting the diversity of multiomics data (Zhang et al., 2019a; Ektefaie et al., 2023). Some existing methods have extended learning to heterogeneous graphs (Zitnik et al., 2024; Ruiz et al., 2023; Xu et al., 2023a), but they often depend on pre-existing knowledge graphs to encode semantic relationships between entities such as genes, patients, and diseases. However, such knowledge graphs are not always available, and constructing them typically requires substantial domain expertise and can be costly. Therefore, we are interested in how to automatically construct heterogeneous graphs by leveraging auxiliary information inherent in multiomics datasets.

Research Question 3: *How can we integrate multiomics data in the presence of different missing-modality patterns, enabling predictions directly from partial modalities?*

The success of graph-based methods for multiomics integration has often relied on the assumption that all omics modalities are available for each patient (Wang et al., 2014, 2021b).

In practice, however, missing modalities, where all data from some modalities are missing for some patients, are an unavoidable challenge in real-world biomedical applications (Mitra et al., 2023). Although different strategies have been proposed to address this issue, many assume that missingness occurs in only a single modality (Hayat et al., 2022; Yao et al., 2024). Moreover, several approaches handle missing modalities only at test time (Tsai et al., 2019; Ma et al., 2021; Reza et al., 2024; Li et al., 2025), overlooking cases where missingness arises in both training and test data. In healthcare, missing modalities can follow different patterns, making this problem even more complex. This raises the question of how to design multiomics integration methods that effectively account for different missing-modality patterns.

Research Question 4: *How can we develop a multimodal feature selection method that explicitly incorporates missing modalities into the selection process?*

Missing modalities are common in multimodal datasets, and different learning strategies have been developed to handle them during model training. However, most feature selection methods still assume complete data and overlook missing modalities when identifying informative features. This gap motivates the need for multimodal feature selection methods that can operate effectively under missing-modality conditions.

1.3 Contributions

This thesis presents four key contributions that collectively address the research questions introduced earlier.

Contribution 1: We introduce `MAgentOmics`, the first general multimodal feature selection framework based on a multi-agent system (MAS) that operates on graph representations of the feature spaces across all omics. Unlike approaches that apply feature selection independently to each modality, this strategy mitigates the curse of dimensionality by modeling both intra- and inter-omics interactions (**Research Question 1**). This is achieved by extending conventional MAS algorithms from single-modality settings to multimodal scenarios through the introduction of inter-omics edges and a collaborative search strategy among agents.

MAgentOmics advances earlier work by enabling agents to share knowledge and iteratively improve their solutions, leading to improved performance.

Contribution 2: We propose HeteroGATomics, a dual-view approach to automatically construct heterogeneous graphs from multiomics data. Each omics modality is represented from two perspectives: (i) a feature similarity network capturing relationships between features, and (ii) a patient similarity network capturing relationships between patients. These two views are combined into a heterogeneous graph with two distinct node types (patients and features) and three relation types ("patient–similar–patient", "feature–similar–feature", and "feature–attribute–patient"). This unified representation captures both patient-level and feature-level relationships, thereby providing a more comprehensive view of the data. By modeling such heterogeneous structures, HeteroGATomics enhances the expressive power of GNNs for multiomics integration (**Research Question 2**). To enable multimodal integration, HeteroGATomics models cross-modality interactions at both the feature level, using the MAS algorithm, and the label level, through a late fusion strategy. Moreover, the multimodal feature selection algorithm not only reduces dimensionality but also enriches the structural information of the heterogeneous graph, benefiting downstream learning.

Contribution 3: We present MAGNET, a missing-modality-aware framework for direct prediction with partial omics data. Specifically, we introduce a patient–modality multi-head attention mechanism that fuses different modalities and enables the model to assess the importance of each modality for individual patients. This mechanism handles diverse missing-modality patterns through a binary modality mask, ensuring that only available modalities contribute to the fused representation (**Research Question 3**). In addition, MAGNET scales linearly with the number of modalities, and its modular design keeps the model both simple and expandable. To maintain consistency of patient representations during fusion, we incorporate a Kullback–Leibler (KL) divergence-based loss that minimizes the misalignment between patient similarity distributions. We further propose a novel patient interaction graph that incorporates missing-modality patterns directly into its structure, improving graph representation learning. Finally, MAGNET can be adapted to incorporate fundamentally

different modalities, supporting broader applicability and extending its potential use beyond biomedical domains.

Contribution 4: We introduce `MMAgentOmics`, a multimodal feature selection framework that explicitly handles missing modalities within the selection process. Rather than excluding samples with missing modalities, `MMAgentOmics` incorporates modality missingness into its state transition mechanism, enabling agents to learn from partial information and select informative features accordingly (**Research Question 4**). We demonstrate its generalizability by applying it to a real biological case study investigating the effects of high-fat diet and metformin on the mouse brain, providing meaningful insights into metabolic and treatment-related molecular responses.

1.4 Thesis Outline

To advance the development of powerful graph-based multiomics integration methods, this thesis introduces several novel frameworks for more accurate cancer classification. This section outlines the structure of the thesis, which is also summarized in Figure 1.1.

In **Chapter 2**, we provide a brief overview of biological concepts related to cancer and the different omics modalities. We then introduce fundamental notions of graphs and deep graph learning models, followed by definitions of multi-agent systems and multimodal learning, which together establish the foundation for the methods developed in this thesis. Finally, we review relevant literature on multimodal fusion for both complete data and settings with missing modalities.

In **Chapter 3**, we present a graph-based multimodal feature selection architecture designed to mitigate the high dimensionality of multiomics data. In this framework, each omics modality is represented as a feature-level graph, where nodes indicate features and edges capture their relationships. The graphs from all modalities are then connected to form a unified search space that incorporates both intra- and inter-omics interactions. To explore this space, we propose an MAS strategy that extends feature selection from single-modality settings to multi-modality scenarios.

In **Chapter 4**, we focus on constructing heterogeneous graphs for multiomics cancer diagnosis. Each omics modality is represented through two views: a patient similarity network capturing patient interactions and a feature similarity network capturing feature interactions. The feature similarity network is enriched using an improved MAS algorithm from Chapter 3, which both reduces feature dimensionality and provides structural information for the heterogeneous graph. Predictions are generated on each heterogeneous graph with a GNN, and late fusion integrates the modality-specific predictions in a supervised manner. This framework enables cross-modality interactions at both the feature and label levels.

In **Chapter 5**, we address the challenge of missing modalities in multiomics cancer classification. We introduce a flexible attention-based framework that fuses available modalities according to their importance and missingness, with model complexity scaling linearly with the number of modalities while adapting to diverse missing patterns. To generate predictions, we construct a patient interaction graph with fused multimodal embeddings as node features and connectivity determined by modality missingness, followed by a conventional GNN. We validate the framework on multiomics datasets with real-world missingness, demonstrating its effectiveness for cancer classification.

In **Chapter 6**, we extend our multimodal feature selection framework to explicitly account for missing modalities. We modify the state transition rules so that feature selection leverages within-omics interactions based on available samples in each modality and cross-omics interactions based on matched samples across modalities. We evaluate this method on a real biological case study in mice to investigate how a high-fat diet and metformin influence brain function, addressing two biological questions: whether diet or metformin directly affects the brain, and what molecular changes occur in the brain under chronic obesity or metformin treatment.

In **Chapter 7**, we conclude by summarizing the findings and outlining promising directions for future work that build upon the contributions of the previous chapters.

1.5 List of Publications

The research presented in this thesis is based primarily on the following publications:

- **Sina Tabakhi** and Haiping Lu, "Multi-agent Feature Selection for Integrative Multi-omics Analysis", 44^{th} *Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1638-1642, 2022.

- **Sina Tabakhi**, Mohammod N. I. Suvon, Pegah Ahadian, and Haiping Lu, "Multimodal Learning for Multi-omics: A Survey", *World Scientific Annual Review of Artificial Intelligence*, 1, 2250004, 2023.

- **Sina Tabakhi**, Charlotte Vandermeulen, Ian Sudbery, and Haiping Lu, "Heterogeneous Graph Attention Network with Joint Feature Selection for Cancer Multiomics Integration," Preprint: `arXiv:2408.02845`, 2024.

- **Sina Tabakhi**, Charlotte Vandermeulen, Ian Sudbery, and Haiping Lu, "Heterogeneous Graph Attention Network Improves Cancer Multiomics Integration," 33^{rd} *Annual International Conference on Intelligent Systems for Molecular Biology (ISMB/ECCB)*, 2025. (**Abstract accepted**).

- **Sina Tabakhi** and Haiping Lu, "Missing-Modality-Aware Graph Neural Network for Cancer Classification," Preprint: `arXiv:2506.22901`, 2025.

- Xianyuan Liu, . . . , **Sina Tabakhi**, . . . , Haiping Lu, "Towards Deployment-Centric Multimodal AI Beyond Vision and Language", *Nature Machine Intelligence*, 7, 1612–1624, 2025.

The following research and publications have also been conducted during the PhD but are not included here, as they are either not directly aligned with the central narrative of the thesis or fall outside its primary scope.

- **Sina Tabakhi** and Parham Moradi, "Universal Feature Selection Tool (UniFeat): An Open-Source Tool for Dimensionality Reduction," *Neurocomputing*, 535, 156-165, 2023.

- **Sina Tabakhi**, "Developing a FAIR-Compliant Open-Source Feature Selection Tool: UniFeat," The University of Sheffield, Report: `shef.data.23702337.v1`, 2023.

- Jiayang Zhang, Xianyuan Liu, Wei Wu, **Sina Tabakhi**, Wenrui Fan, Shuo Zhou, Kang Lan Tee, Tuck Seng Wong, and Haiping Lu, "Classifying the Stoichiometry of Virus-Like Particles with Interpretable Machine Learning", 47^{th} *Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2025.

- Sayedmohammadreza Rastegari, **Sina Tabakhi**, Xianyuan Liu, Wei Sang, and Haiping Lu, "Co-evolution-Based Metal-Binding Residue Prediction with Graph Neural Networks", 33^{rd} *Annual International Conference on Intelligent Systems for Molecular Biology (ISMB/ECCB)*, 2025. (**Abstract accepted**).

- Xiaokun Liu, . . . , **Sina Tabakhi**, . . . , Haiping Lu, "Interpretable Multimodal Learning for Tumor Protein–Metal Binding: Progress, Challenges, and Perspectives", *Methods*, 242, 97-112, 2025.

- Prasun C. Tripathi, **Sina Tabakhi**, Mohammod N. I. Suvon, Lawrence Schöb, Samer Alabed, Andrew J. Swift, Shuo Zhou, and Haiping Lu, "Interpretable Multimodal Learning for Cardiovascular Hemodynamics Assessment," *IEEE Transactions on Medical Imaging*, 2026. (**Accepted**).

Chapter 2

Background

This thesis focuses on the design of multimodal learning methods with graphs for multiomics integration. Accordingly, this chapter introduces the necessary biological and computational background relevant to this work. Section 2.1 covers the biological foundation, introducing cancer biology (Section 2.1.1) and multiomics data (Section 2.1.2). Section 2.2 presents the computational background, beginning with fundamental concepts of graphs and graph types (Section 2.2.1), followed by graph neural networks (Section 2.2.2), relational graph neural networks (Section 2.2.3), multi-agent systems (Section 2.2.4), and multimodal learning (Section 2.2.5). Finally, we review different multimodal fusion strategies for complete data (Section 2.3) and for data with missing modalities (Section 2.4).

2.1 Biological Background

2.1.1 Cancer Biology

Cancer is one of the leading causes of death worldwide, with nearly 10 million deaths in 2020, representing approximately one in six of all deaths. The most commonly diagnosed types are breast, lung, colorectal, and prostate cancers (Pettini et al., 2021). Cancer refers to a group of diseases characterized by abnormal and uncontrolled cell growth, resulting from changes in cell division and cell death (Brown et al., 2023; Ruddon, 2007). These cellular

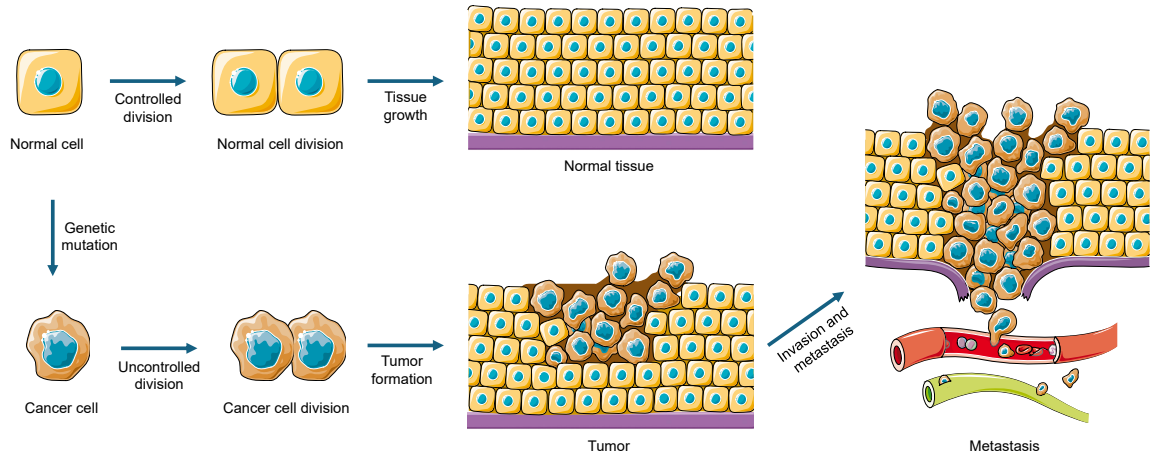

Fig. 2.1 **Cancer development and progression.** Normal cell division supports healthy tissue growth. Cancer originates from a single abnormal cell, where genetic mutations can lead to uncontrolled cell division, tumor formation, and metastasis. Icon credit: metastasis icon by Servier is licensed under CC-BY 3.0, with additional icons adapted from the original.

changes arise from underlying genetic and molecular alterations, in which the DNA of cells is altered, disrupting normal regulatory mechanisms and providing incorrect instructions for cellular functioning. Such mutations drive the development of cancer cells and are observed across all cancer types. If left untreated, abnormal cell growth can invade nearby tissues and metastasize to distant sites, which is the main cause of cancer-related death (Hejmadi, 2014). Figure 2.1 shows the stages of cancer development and progression, from normal cell division to metastasis.

The progression from a normal cell to a malignant phenotype is governed by the acquisition of specific biological capabilities. These capabilities are commonly described as the hallmarks of cancer, which provide a conceptual framework for understanding the complexity of human cancers. Initially, six hallmarks were defined, including sustained proliferative signalling, evasion of growth suppressors, resistance to cell death, replicative immortality, angiogenesis, and activation of invasion and metastasis (Hanahan and Weinberg, 2000). Subsequent studies have expanded this framework to include additional capabilities, such as deregulated cellular energetics, evasion of immune destruction, and phenotypic plasticity, further contributing to tumor development (Hanahan and Weinberg, 2011; Hanahan,

2026). Collectively, these hallmarks reflect disruptions in fundamental cellular processes and regulatory mechanisms, providing a functional basis for cancer progression.

At the molecular level, these hallmarks are primarily driven by alterations in two major classes of genes: proto-oncogenes and tumor suppressor genes (Hanahan and Weinberg, 2011; Prasad and Sitara, 2026). Proto-oncogenes normally regulate cell growth and signalling; however, when activated through mutation or overexpression, they become oncogenes that promote uncontrolled cellular proliferation through gain-of-function effects (Kontomanolis et al., 2020). In contrast, tumor suppressor genes regulate cell cycle progression, DNA repair, and apoptosis to maintain genomic stability. Loss-of-function alterations or epigenetic silencing of tumor suppressor genes remove critical regulatory checkpoints, thereby facilitating tumor development (Lee and Muller, 2010).

These functional changes arise through a diverse range of molecular alterations affecting the genome and its regulation, among which genetic mutations play a central role in cancer initiation and progression. A mutation refers to an alteration in the genomic sequence. These can be broadly divided into driver and passenger mutations (Aparisi et al., 2019; Kumar et al., 2020). Driver mutations lead to tumor initiation and growth with a selective advantage, whereas passenger mutations accumulate during tumor evolution without directly contributing to cancer development. Although drivers are the primary focus of cancer sequencing efforts, passenger mutations also provide valuable insights into mutational processes and evolutionary history, and therefore require systematic study. Most mutations in the genome have no immediate effect, but when they occur in important regulatory regions or genes, they can disrupt cellular processes. For example, a mutation may alter the mRNA sequence of a gene, producing an abnormal protein that no longer functions correctly (Schulte-Sasse, 2020). While this illustrates how mutations can drive downstream alterations, cancer progression is also shaped by independent molecular changes at the RNA, protein, and epigenetic levels, which influence tumor metabolism, signaling pathways, and metastatic potential (Marei, 2025; Orsolic et al., 2023).

The development and progression of cancer from a nonmalignant to a malignant state is a dynamic and heterogeneous process. Tumors show variability both across patients

(inter-tumor heterogeneity) and within a single tumor (intra-tumor heterogeneity) (Burrell et al., 2013). Inter-tumor heterogeneity refers to differences in genetic and molecular profiles between patients with the same cancer type, whereas intra-tumor heterogeneity arises from the presence of multiple subclonal populations within a tumor (Dagogo-Jack and Shaw, 2018). This diversity is driven by the continuous accumulation of genetic and epigenetic alterations, together with selective pressures from the tumor microenvironment and therapeutic interventions. As tumors evolve, distinct cellular subpopulations may acquire different functional properties, such as increased invasiveness or resistance to treatment (McGranahan and Swanton, 2017). Consequently, cancer can be viewed as a dynamic and evolving system rather than a static disease.

Capturing the complexity and heterogeneity of cancer has been enabled by advances in next-generation sequencing technologies, which enable comprehensive profiling across multiple molecular layers, including the genome, epigenome, transcriptome, proteome, and metabolome. These technologies provide a multi-dimensional view of tumor biology, allowing the identification of diverse molecular alterations underlying cancer development. Large-scale initiatives such as TCGA have generated extensive datasets, revealing thousands of tumor-associated mutations and facilitating the discovery of oncogenic drivers and candidate therapeutic targets.

The generation of large-scale multiomics datasets has transformed the prospects for precision medicine in oncology. Precision medicine aims to tailor prevention, diagnosis, and treatment strategies to the molecular characteristics of an individual patient's tumor. Cancer biomarkers are measurable molecular signatures associated with disease presence, subtype, or progression and are central to precision oncology. For example, the identification of BRCA1/2 mutations in breast and ovarian cancers informs the use of PARP inhibitors (Dibitetto et al., 2024), while alterations in EGFR or ALK guide targeted therapies in lung cancer (Bouchard and Daaboul, 2025). Beyond single-gene markers, integrative analyses of multiomics data are increasingly being used to define molecular subtypes, predict treatment response, and identify novel therapeutic targets. However, the complexity and heterogeneity of cancer highlight the need for computational approaches that can effectively integrate

diverse data modalities to uncover actionable biomarkers and improve clinical decision-making (Perez-Lopez et al., 2024).

2.1.2 Multiomics Data

Multiomics data are derived from different sources, including molecular and non-molecular omics modalities. Molecular omics represent a major dimension of multiomics data, aiming to characterize the molecular biology of human diseases through genomics, transcriptomics, proteomics, metabolomics, epigenomics, microbiomics, and exposomics. Complementary perspectives are provided by non-molecular modalities, specifically radiomics (Lambin et al., 2012) and phenomics (Bilder et al., 2009), which incorporate imaging and clinical information into the multiomics landscape. Each of these omics modalities has been studied independently to address biomedical questions. While different sources provide complementary perspectives and reveal unique aspects of diseases, integrating multiple omics modalities has become increasingly common in existing research. Introducing the distinct modalities of omics data and their potential applications highlights the breadth of opportunities available for addressing biomedical challenges (Karczewski and Snyder, 2018; Acosta et al., 2022). Table 2.1 summarizes the main modalities of omics data and their applications, while Figure 2.2 illustrates how these data sources and their fusion contribute to various biomedical tasks.

We review multiomics data in the following subsections, dividing them into molecular and non-molecular omics based on their prevalence in research.

Molecular omics modalities

Genes, located within the deoxyribonucleic acid (DNA) of human cells, encode proteins that are essential for human development and physiological maintenance. Alterations in genes and their regulation contribute to many life-threatening diseases, including cancer, making the study of genetic and molecular constituents important for diagnosis and treatment.

The advent of next-generation sequencing (NGS) has enabled the large-scale characterization of molecular data, collectively referred to as omics (Karczewski and Snyder,

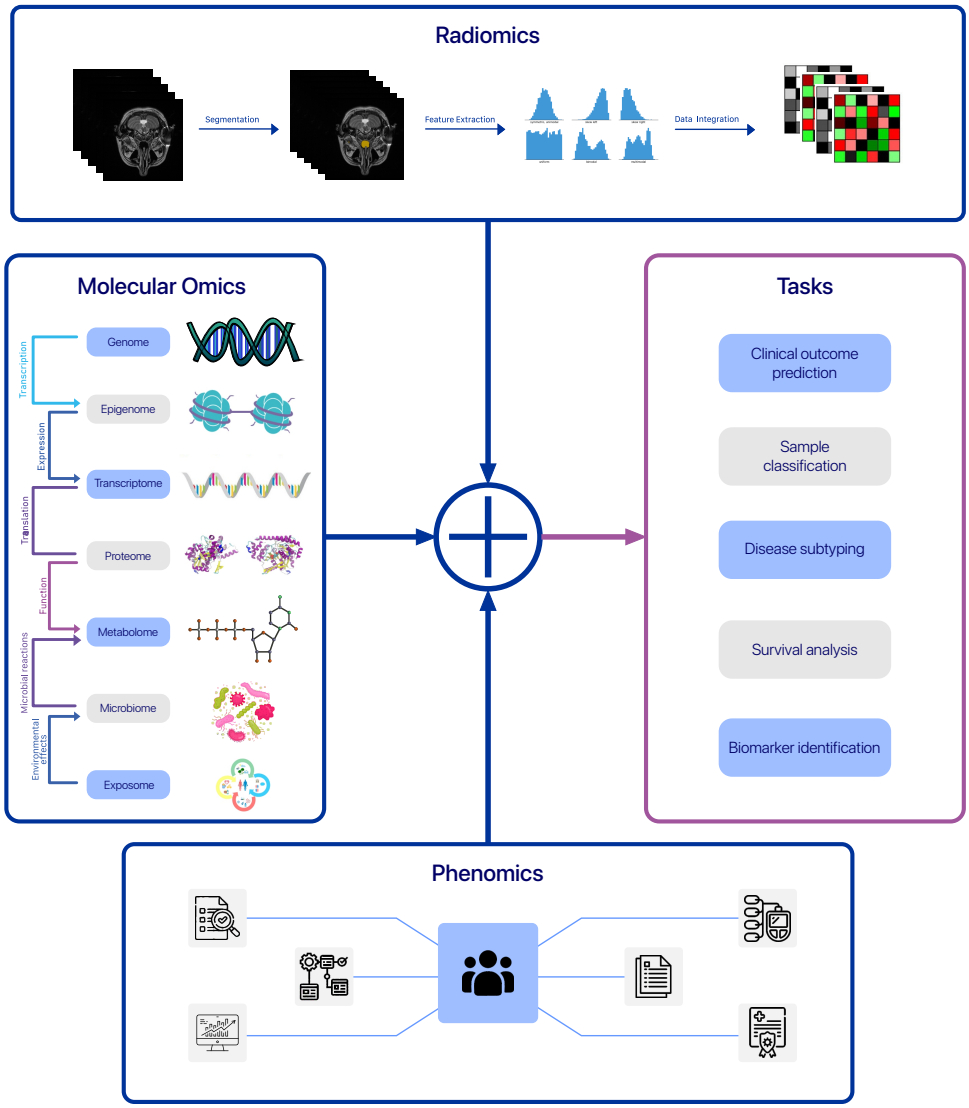

Fig. 2.2 **Multiple omics modalities and their fusion for various tasks.**

2018; da Fonseca et al., 2016). This development allows researchers to study molecular layers, including genomics, transcriptomics, proteomics, metabolomics, and epigenomics. Integrating these omics modalities provides insights into gene expression, regulation, and protein activity, supporting several biomedical tasks, including disease diagnosis, prognosis, treatment, prevention, and biomarker identification.

The complete set of genetic information required for human development and growth is contained in the genome, with its instructions expressed through ribonucleic acid (RNA).

Table 2.1 Overview of multiomics data and their applications in diagnosis, prognosis, and precision medicine.

Omics modality	Description	Types/Technologies	Biomedical applications
Genomics	Genomics studies the structure, function, and sequencing of DNA to identify genetic variants associated with disease, treatment response, and prognosis.	Structural genomics (Skolnick et al., 2000), Functional genomics (Cano-Gamez and Trynka, 2020), Comparative genomics (Hardison, 2003), Mutation genomics (Stenson et al., 2009)	Gene discovery and diagnosis of rare genetic disorders (e.g., heart disease and cancers); Identification and diagnosis of genetic factors contributing to prevalent diseases; Pharmacogenetics and targeted therapy; Diagnosis of infectious diseases
Epigenomics	Epigenomics studies heritable changes in gene expression that occur without alterations in the DNA sequence, playing a key role in development and homeostasis.	DNA methylation (Singal and Ginder, 1999), Histone modification (Karlić et al., 2010), Non-coding RNA (Eddy, 2001)	Early-stage detection, e.g., coronary artery disease; Therapeutic development
Transcriptomics	Transcriptomics examines the transcriptome, the complete set of RNA transcripts in a cell, to understand gene function and regulation.	DNA microarrays (Stoughton, 2005), RNAseq (Hrdlickova et al., 2017)	Identification of early cancer biomarkers; Analysis of human and pathogen transcriptomes; Response to the environment; Gene function annotation; Non-coding RNA
Proteomics	Proteomics identifies and quantifies the full set of proteins in a cell, tissue, or organism.	Expression proteomics (Banks et al., 2000), Functional proteomics (Colland et al., 2004), Structural proteomics (Renfrey and Featherstone, 2002)	Protein production, degradation, and steady-state abundance rates; Protein movement between subcellular compartments; Protein involvement in metabolic pathways; Protein interaction display used in the drug discovery process
Metabolomics	Metabolomics analyzes small-molecule metabolites, the substrates and products of metabolism, which are shaped by genetic and environmental factors.	Metabolite fingerprinting (Krishnan et al., 2005), Metabolic profiling, Targeted analysis	Investigation of several human diseases (e.g., cancers and natural metabolic errors); Design improved therapeutic strategies; Toxicology (i.e., toxicological effects); Pharmacology (i.e., nutrition)
Microbiomics	Microbiomics studies the composition and function of microbial communities in the human body and their roles in health and disease.	Microbiome (Schwabe and Jobin, 2013)	Infectious disease diagnosis; Monitoring microbial components in chronic diseases (e.g., heart disease, diabetes, and chronic lung disease)
Exposomics	Exposomics investigates the cumulative effects of environmental exposures on health and disease development.	Environmental exposures (Price et al., 2022)	New insights into the development of chronic diseases; Reveal nongenetic disease causes
Radiomics	Radiomics extracts quantitative features from medical images using computational algorithms to reveal patterns and characteristics of tissues and diseases.	Medical images	Diagnostic differentiation of suspected tissue; Survival prognosis; Prediction of clinical responses; Prediction risk of distant metastasis (Haider et al., 2020)
Phenomics	Phenomics investigates variations in human phenotypes arising from gene–environment interactions, supporting personalized medicine and biomedical research.	Qualitative traits (Lanktree et al., 2010), Quantitative traits	Functional genomics; Pharmaceutical research; Disease risk prediction

Building on this foundation, different omics disciplines investigate molecular layers of biological systems at varying levels. *Genomics* focuses on the study of the genome, aiming to identify variations in DNA sequences that contribute to human diseases. *Transcriptomics* examines gene expression by analyzing RNA transcripts, identifying changes that reveal how diseases influence cellular and organismal functions. *Proteomics* investigates the complete set of proteins in a biological system, complementing information from genomics and often relying on mass spectrometry (MS) (Keller et al., 2005). *Metabolomics* profiles cellular metabolites and their interactions, providing a fingerprint of cellular physiology that supports the detection of metabolic disorders and biomarker discovery. *Epigenomics* explores heritable modifications in gene regulation that occur without changes in the DNA sequence, offering insights into how environmental factors influence disease. *Microbiomics* studies microbial communities, including bacteria, archaea, fungi, and viruses, to understand their roles in human health and disease. *Exposomics* measures the total environmental exposures experienced throughout an individual's life to assess their effects on health and the complex interaction between genetics and environment.

Several additional omics modalities have been developed but are not yet routinely incorporated into molecular omics analyses. For example, *metagenomics*, *metatranscriptomics*, and *metaproteomics* extend foundational omics modalities to complex microbial communities, but their application still faces major challenges, including limitations of sequencing technologies (short read lengths and high error rates) as well as economic constraints (Aguiar-Pulido et al., 2016). Similarly, *glycomics*, which provides a comprehensive view of glycans synthesized in cells, has yet to be fully integrated with other omics data types (Kunej, 2019).

Non-molecular omics modalities

With advances in medical imaging technologies and the growing recognition of the importance of clinical data, complementary omics modalities, radiomics and phenomics, have emerged to provide additional insights into biological and clinical problems.

In precision medicine, *radiomics* extracts quantitative features from medical images to support decision-making systems for accurate diagnosis (Shur et al., 2021). Its workflow

typically involves image acquisition from modalities such as computed tomography (CT), magnetic resonance imaging (MRI), or digital histopathology, followed by segmentation to identify regions of interest. Quantitative features describing intensity, shape, and texture are then extracted and organized into structured datasets, which are subsequently analyzed using machine learning models (Caruso et al., 2021). Access to standardized image collections is a critical requirement for this process. A major initiative in this direction is The Cancer Imaging Archive (TCIA), which provides publicly accessible imaging data for a variety of cancers, including lung cancer (Clark et al., 2013).

Another complementary modality is *phenomics*, which systematically measures the full set of qualitative and quantitative phenomes, including physical and biochemical traits. It examines how environmental and lifestyle factors interact with genetic variation to influence disease risk (Houle et al., 2010). A key area is electronic health record (EHR)-based phenotyping, which extracts clinical characteristics from patient records to advance our understanding of health and disease (Richesson et al., 2013). By integrating genetic and phenotypic data, phenomics offers unique insights into the genotype–phenotype relationship underlying complex human diseases (Bilder et al., 2009).

2.2 Computational Background

2.2.1 Graphs

Graphs – or networks – are data structures that provide a universal way to represent interactions among entities in complex systems. In general, a graph consists of a set of entities or objects (i.e., nodes) together with the relationships between pairs of entities (i.e., edges). This representation enables researchers to capture and analyze these interactions. Many real-world systems can be naturally modeled as graphs, such as molecular interaction networks in biology or healthcare systems in medicine (Li et al., 2022a; Ektefaie et al., 2023).

From a mathematical perspective, a graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of N nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of edges. Each edge $(u, v) \in \mathcal{E}$ connects a node $u \in \mathcal{V}$ to a node $v \in \mathcal{V}$. A graph may also contain *self-loops*, where a node is connected to itself, i.e.,

$(u,u) \in \mathcal{E}$. The *neighborhood* of a node $u \in \mathcal{V}$ is the set of nodes directly connected to u, defined as $\mathcal{N}(u) = \{ v \in \mathcal{V} \mid (u,v) \in \mathcal{E} \}$.

The graph $\mathcal{G}$ is *directed* if the order of nodes in an edge matters, i.e., $(u,v) \neq (v,u)$, representing an asymmetric relationship. Conversely, $\mathcal{G}$ is *undirected* if edges are unordered pairs, i.e., $(u,v) = (v,u)$, representing symmetric relationships. A common way to represent a graph $\mathcal{G}$ is through its *adjacency matrix* $\mathbf{A} \in \{0,1\}^{N \times N}$, where the entry a_{ij} indicates whether there is an edge between nodes i and j. Specifically,

$$a_{ij} = \begin{cases} 1, & \text{if } (i,j) \in \mathcal{E}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

Based on this definition, a graph $\mathcal{G}$ is *unweighted* if its adjacency matrix $\mathbf{A} \in \{0,1\}^{N \times N}$ has binary entries, with $a_{ij} = 1$ indicating the presence of an edge and $a_{ij} = 0$ otherwise. It is *weighted* if $\mathbf{A} \in \mathbb{R}^{N \times N}$, where $a_{ij} = w_{ij}$ denotes the weight of the edge between nodes i and j. A graph is called *sparse* if the number of edges is significantly smaller than the maximum possible number of edges for a given number of nodes.

To represent important characteristics of nodes and edges, we can associate feature vectors with them. For example, in a biomedical setting, a node representing a patient may have associated biological information. Node features for all nodes are collected in a *node feature matrix* $\mathbf{X} \in \mathbb{R}^{N \times D}$, where each row corresponds to a node and D is the feature dimension. Similarly, edge features for all edges are collected in an *edge feature matrix* $\mathbf{E} \in \mathbb{R}^{|\mathcal{E}| \times F}$, where each row corresponds to an edge and F is the feature dimension. For each node, we write the feature vector as $\mathbf{x}_i \in \mathbb{R}^D, i = 1, \ldots, N$. For each edge, we write the feature vector as $\mathbf{e}_{ij} \in \mathbb{R}^F, (i,j) \in \mathcal{E}$.

Homogeneous graphs

A graph is called *homogeneous* if it consists of a single type of node and a single type of edge. The general definitions introduced above – nodes, edges, adjacency matrices, and node/edge features – can be directly applied to homogeneous graphs, since all entities and relationships are of the same kind.

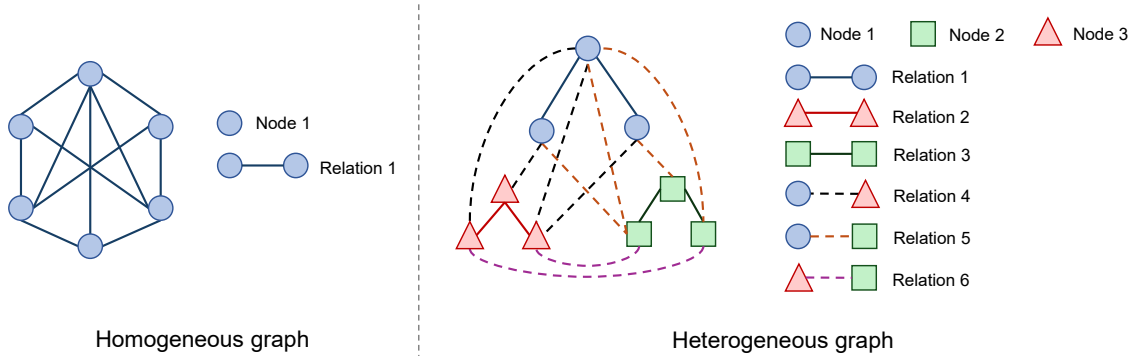

Fig. 2.3 **Illustration of homogeneous and heterogeneous graphs.** A homogeneous graph (left) contains a single type of node and relation, whereas a heterogeneous graph (right) consists of multiple node types and relation types.

Heterogeneous graphs

Many real-world systems involve multiple types of entities and relationships. A *heterogeneous* graph extends the homogeneous setting by allowing different types of nodes and/or edges. For example, in biomedical applications, nodes may represent patients, genes, or clinical features, while edges may encode distinct relationships such as patient–patient similarity, gene–gene interactions, or patient–feature associations. Formally, a heterogeneous graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \phi, \psi)$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges. Each node is associated with a type via a mapping function $\phi : \mathcal{V} \to \mathcal{A}$, where $\mathcal{A}$ is the set of node types. Similarly, each edge is associated with a type via a mapping function $\psi : \mathcal{E} \to \mathcal{R}$, where $\mathcal{R}$ is the set of edge types or relations (Hamilton, 2020; Luo et al., 2022). The graph is heterogeneous if either $|\mathcal{A}| > 1$ or $|\mathcal{R}| > 1$. Figure 2.3 provides an example of a homogeneous and a heterogeneous graph.

2.2.2 Graph Neural Networks

With the success of deep learning approaches such as convolutional neural networks (CNNs) in the image domain (Li et al., 2021b; Gu et al., 2018) and recurrent or transformer-based models in the text domain (Vaswani et al., 2017; Xu et al., 2023b), their generalization to graphs has been increasingly studied. Deep learning on graph-structured data, known as graph neural networks (GNNs), has become an active research area in systems biology and

healthcare (Li et al., 2022a; Ektefaie et al., 2023; Wang et al., 2021b; Ma et al., 2025). The objective of GNNs is to learn node representations (embeddings) from the graph structure and, when available, from associated feature information on nodes and edges, in order to support downstream tasks (Acosta et al., 2022; Hamilton et al., 2018). Among the various tasks defined on graphs, this thesis focuses on supervised *node-level prediction*. In this setting, each node $u \in \mathcal{V}$ is associated with a label $\mathbf{y}_u \in \mathbf{Y}$, and the goal is to learn an embedding $\mathbf{h}_u$ such that it can be used to accurately predict $\mathbf{y}_u$. The label matrix $\mathbf{Y} \in \mathbb{R}^{N \times C}$ contains one-hot encoded labels for N patients across C classes.

To perform node-level prediction with GNNs, we rely on the neural message-passing scheme (Gilmer et al., 2017), which forms the foundation of modern GNNs. This scheme is based on exchanging messages between neighboring nodes, followed by aggregation and update steps that refine each node's representation using neural networks. In other words, during each message-passing iteration, the representation $\mathbf{h}_u$ of a node $u \in \mathcal{V}$ is updated by aggregating information from its neighbors $\mathcal{N}(u)$. Let $\mathbf{h}_u^{(l)}$ denote the hidden representation of node u at layer l; then, a generic GNN layer updates node embeddings as:

$$\mathbf{m}_{u \leftarrow v}^{(l)} = \textsc{Message}^{(l)}\big(\mathbf{h}_u^{(l)}, \mathbf{h}_v^{(l)}, \mathbf{e}_{uv}\big), \tag{2.2}$$

$$\mathbf{m}_{\mathcal{N}(u)}^{(l)} = \textsc{Aggregate}^{(l)}\big(\{\mathbf{m}_{u \leftarrow v}^{(l)} : v \in \mathcal{N}(u)\}\big), \tag{2.3}$$

$$\mathbf{h}_u^{(l+1)} = \textsc{Update}^{(l)}\big(\mathbf{h}_u^{(l)}, \mathbf{m}_{\mathcal{N}(u)}^{(l)}\big). \tag{2.4}$$

This framework consists of three functions: $\textsc{Message}(\cdot)$, $\textsc{Aggregate}(\cdot)$, and $\textsc{Update}(\cdot,\cdot)$. In the message step (Equation (2.2)), each node generates a message using the $\textsc{Message}(\cdot)$ function, which is sent to its neighboring nodes. The message can incorporate information about the edge between nodes u and v. When edge features $\mathbf{e}_{uv}$ are available, they are typically concatenated with the source node representation $\mathbf{h}_v^{(l)}$ before transmission, and the message function can be implemented as a simple linear layer that transforms this concatenated vector.

The $\textsc{Aggregate}(\cdot)$ and $\textsc{Update}(\cdot,\cdot)$ functions are arbitrary differentiable and typically parameterized functions, often implemented using neural networks. In the aggregation step (Equation (2.3)), all messages from the neighborhood $\mathcal{N}(u)$ of node u are combined

using the $\text{AGGREGATE}(\cdot)$ function to produce a single message vector that summarizes the neighborhood information. The aggregation function must be permutation invariant to the order of neighbors. Common choices include summation, mean, or max pooling. In the update step (Equation (2.4)), the aggregated neighborhood message is combined with the node's previous representation $\mathbf{h}_u^{(l)}$ through the $\text{UPDATE}(\cdot,\cdot)$ function, resulting in the updated node representation $\mathbf{h}_u^{(l+1)}$. A common choice is to combine the aggregated neighborhood message with the node's previous representation either by a weighted sum (Kipf and Welling, 2017) or by concatenation (Hamilton et al., 2017), typically followed by a nonlinear transformation.

A GNN starts with initial node representations at layer $l = 0$, which are set to the input node features, i.e., $\mathbf{h}_u^{(0)} = \mathbf{x}_u, \forall u \in \mathcal{V}$. After running L layers of message passing, the output of the final layer is taken as the learned representation of each node, i.e., $\mathbf{z}_u = \mathbf{h}_u^{(L)}, \forall u \in \mathcal{V}$. Different implementations of the message, aggregation, and update steps have led to the development of many GNN models (Veličković et al., 2018; Xu et al., 2019; Hamilton et al., 2017; Kipf and Welling, 2017; Zhang et al., 2018c). In the following, we introduce three representative and concrete examples.

Graph convolutional networks

Graph convolutional networks (GCNs) (Kipf and Welling, 2017) are one of the most widely used GNN models and can be viewed as a special case of the general message-passing framework. Intuitively, GCNs extend the convolution operation used in image processing to graphs, by aggregating and updating feature information between neighboring nodes while applying symmetric degree normalization to avoid scaling issues. Formally, a GCN layer defines the message-passing scheme as:

$$\mathbf{m}_{u\leftarrow v}^{(l)} = \frac{1}{\sqrt{|\mathcal{N}(u)||\mathcal{N}(v)|}} \mathbf{W}^{(l)} \mathbf{h}_v^{(l)}, \tag{2.5}$$

$$\mathbf{m}_{\mathcal{N}(u)}^{(l)} = \sum_{v \in \mathcal{N}(u) \cup \{u\}} \mathbf{m}_{u\leftarrow v}^{(l)}, \tag{2.6}$$

$$\mathbf{h}_u^{(l+1)} = \textsc{ReLU}\big(\mathbf{m}_{\mathcal{N}(u)}^{(l)}\big), \tag{2.7}$$

where $\mathbf{W}^{(l)}$ is a learnable weight matrix at layer l, $|\mathcal{N}(u)|$ denotes the degree of node u, and $\textsc{ReLU}(\cdot)$ is the nonlinear activation function. In GCNs, self-loops are included as part of the neighborhood aggregation in Equation (2.6).

GraphSAGE

Graph sample and aggregation (GraphSAGE) (Hamilton et al., 2017) extends the message-passing framework by introducing a sampling strategy to handle large graphs. Instead of aggregating information from all neighbors, GraphSAGE samples a fixed-size set of neighbors for each node, making it scalable to graphs with high-degree nodes. Furthermore, GraphSAGE explicitly combines the node's own representation with the aggregated neighborhood information via concatenation, followed by a learnable transformation. Formally, a GraphSAGE layer can be defined as:

$$\mathbf{m}_{u\leftarrow v}^{(l)} = \mathbf{h}_v^{(l)}, \tag{2.8}$$

$$\mathbf{h}_u^{(l+1)} = \sigma\Big(\mathbf{W}^{(l)} \cdot \textsc{Concat}\big(\mathbf{h}_u^{(l)}, \mathbf{m}_{\mathcal{N}(u)}^{(l)}\big)\Big), \tag{2.9}$$

where $\textsc{Concat}(\cdot,\cdot)$ denotes concatenation of representations and $\sigma(\cdot)$ is a nonlinear activation function. In Equation (2.9), $\mathbf{m}_{\mathcal{N}(u)}^{(l)}$ is defined similarly to Equation (2.6), except that the neighborhood is obtained by uniformly sampling a fixed-size set of neighbors, instead of using the full neighborhood. The $\textsc{Aggregate}(\cdot)$ function in GraphSAGE can be implemented in different ways, such as taking the mean of neighbor embeddings, applying a nonlinear transformation followed by a symmetric function like max pooling, or using an LSTM applied to a randomly ordered sequence of neighbor embeddings (Hamilton et al., 2017).

Graph attention networks

Graph attention networks (GATs) (Veličković et al., 2018) extend the message-passing framework by introducing an attention mechanism (Bahdanau et al., 2015) to learn the importance of different neighbors when aggregating information. Unlike GCNs and GraphSAGE, which apply a fixed normalization based on node degrees, GATs compute attention coefficients that adaptively weight the contribution of each neighbor, allowing the model to focus on the most relevant connections. Formally, a GAT layer can be defined as:

$$\mathbf{m}_{u\leftarrow v}^{(l)} = \mathbf{W}^{(l)}\mathbf{h}_v^{(l)}, \tag{2.10}$$

$$\mathbf{m}_{\mathcal{N}(u)}^{(l)} = \sum_{v\in\mathcal{N}(u)} \alpha_{uv}^{(l)}\, \mathbf{m}_{u\leftarrow v}^{(l)}, \tag{2.11}$$

$$\mathbf{h}_u^{(l+1)} = \sigma\big(\mathbf{m}_{\mathcal{N}(u)}^{(l)}\big), \tag{2.12}$$

where

$$\alpha_{uv}^{(l)} = \frac{\exp(e_{uv}^{(l)})}{\sum_{k\in\mathcal{N}(u)}\exp(e_{uk}^{(l)})}, \tag{2.13}$$

$$e_{uv}^{(l)} = \textsc{LeakyReLU}\big(\mathbf{a}^{(l)^\top}\cdot \textsc{Concat}(\mathbf{m}_{u\leftarrow u}^{(l)}, \mathbf{m}_{u\leftarrow v}^{(l)})\big), \tag{2.14}$$

$e_{uv}^{(l)}$ is the unnormalized attention score measuring the importance of source node v to destination node u, $\mathbf{a}^{(l)}$ is a trainable attention vector, $(\cdot)^\top$ indicates transposition, and $\exp(\cdot)$ is the standard exponential function. Equation (2.13) applies a softmax operation over the neighborhood $\mathcal{N}(u)$ to normalize these scores into attention coefficients, ensuring that $\sum_{v\in\mathcal{N}(u)}\alpha_{uv}^{(l)} = 1$.

To stabilize the learning process of the attention mechanism, multi-head attention (Veličković et al., 2018) is widely utilized. In this scheme, the outputs of K independent attention heads are combined to form the final node representations, defined as:

$$\mathbf{m}_{u\leftarrow v}^{(l,k)} = \mathbf{W}^{(l,k)}\mathbf{h}_v^{(l)}, \tag{2.15}$$

$$\mathbf{m}_{\mathcal{N}(u)}^{(l,k)} = \sum_{v\in\mathcal{N}(u)} \alpha_{uv}^{(l,k)}\, \mathbf{m}_{u\leftarrow v}^{(l,k)}, \tag{2.16}$$

$$\mathbf{h}_u^{(l+1)} = \textsc{Concat}\Big(\sigma\big(\mathbf{m}_{\mathcal{N}(u)}^{(l,1)}\big), \sigma\big(\mathbf{m}_{\mathcal{N}(u)}^{(l,2)}\big), \ldots, \sigma\big(\mathbf{m}_{\mathcal{N}(u)}^{(l,K)}\big)\Big), \tag{2.17}$$

where $\mathbf{m}_{\mathcal{N}(u)}^{(l,k)}$ is the aggregated message from the neighborhood of node u computed by the k-th attention head, each with its own parameters. If we perform multi-head attention on the final layer of the network, instead of concatenation we apply averaging, as follows:

$$\mathbf{h}_u^{(l+1)} = \frac{1}{K}\sum_{k=1}^{K} \sigma\big(\mathbf{m}_{\mathcal{N}(u)}^{(l,k)}\big). \tag{2.18}$$

2.2.3 Relational Graph Neural Networks

The models discussed in Section 2.2.2 implicitly assume homogeneous graphs. However, in many real-world applications, such as healthcare, the underlying data naturally forms heterogeneous graphs with multiple types of relations. In this section, we review graph neural network strategies that have been specifically designed to handle such heterogeneous and multi-relational graphs.

Relational graph convolutional networks

Relational graph convolutional networks (RGCNs) (Schlichtkrull et al., 2018) are among the first models designed to handle heterogeneous graphs with multiple relation types. They extend GCNs by modifying the aggregation function to incorporate relation-specific transformations. Given a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \phi, \psi)$ with a set of relations $\mathcal{R}$, an R-GCN layer is defined as:

$$\mathbf{m}_{r,u\leftarrow v}^{(l)} = \frac{1}{c_{u,r}} \mathbf{W}_r^{(l)} \mathbf{h}_v^{(l)}, \quad (u,v) \in \mathcal{E}_r, \tag{2.19}$$

$$\mathbf{m}_{\mathcal{N}(u)}^{(l)} = \sum_{r\in\mathcal{R}} \sum_{v\in\mathcal{N}_r(u)} \mathbf{m}_{r,u\leftarrow v}^{(l)}, \tag{2.20}$$

$$\mathbf{h}_u^{(l+1)} = \sigma\Big(\textsc{Sum}\big(\mathbf{m}_{\mathcal{N}(u)}^{(l)}, \mathbf{m}_{u\leftarrow u}^{(l)}\big)\Big). \tag{2.21}$$

where $\textsc{Sum}(\cdot,\cdot)$ denotes element-wise addition of its input vectors, $\mathbf{m}_{u\leftarrow u}^{(l)} = \mathbf{W}_0^{(l)}\mathbf{h}_u^{(l)}$ is the self-loop message with transformation weight matrix $\mathbf{W}_0^{(l)}$, $\mathcal{N}_r(u)$ denotes the neighbors of

node u under relation r, $\mathbf{W}_r^{(l)}$ is a relation-specific weight matrix, and $c_{u,r}$ is a normalization constant (e.g., $c_{u,r} = |\mathcal{N}_r(u)|$).

As RGCNs assign a trainable weight matrix to each relation type, the number of parameters grows rapidly with the number of relations. To address this scalability issue, Schlichtkrull et al. (2018) proposed two regularization strategies: *basis decomposition* and *block-diagonal decomposition* of the relation-specific weight matrices.

Relational graph attention networks

Relational graph attention networks (RGATs) (Chen et al., 2021a) extend the GAT framework to heterogeneous and multi-relational graphs. While standard GATs learn attention coefficients to adaptively weight the importance of neighbors, RGATs incorporate relation types into the attention mechanism. This allows the model to differentiate how information from a neighbor should be weighted depending not only on the neighbor's features but also on the type of relation connecting the two nodes. Formally, relation-aware attention can be expressed by making the attention score relation-specific:

$$e_{r,uv}^{(l)} = \textsc{LeakyReLU}\big(\mathbf{a}_r^{(l)^\top} \cdot \textsc{Concat}(\mathbf{m}_{r,u\leftarrow u}^{(l)}, \mathbf{m}_{r,u\leftarrow v}^{(l)})\big), \tag{2.22}$$

where $r \in \mathcal{R}$ denotes the relation type, and the normalized attention coefficients $\alpha_{r,uv}^{(l)}$ are obtained by applying a softmax function across neighbors under relation r.

By integrating relation information into the attention mechanism, RGATs provide greater flexibility in modeling heterogeneous graphs compared to RGCNs, as they can learn both relation-specific transformations and adaptive neighbor weighting.

2.2.4 Multi-Agent Systems

Multi-agent systems (MAS) (Wooldridge, 2009) are computational frameworks in which multiple simple agents interact and cooperate within a shared environment to solve complex problems that are difficult for a single agent to address. Each agent operates based on local information and simple rules, but through interaction, the collective system can achieve

powerful problem-solving behavior. MAS approaches have been applied in optimization (Goldman and Zilberstein, 2003), robotics (Ota, 2006), distributed control (Wernstedt, 2005), and biological modeling (Roche et al., 2008).

In this thesis, we focus on the application of MAS to feature selection. Specifically, we address high-dimensional multimodal data, where feature selection is a critical step to reduce dimensionality, enhance interpretability, and improve predictive performance. Feature selection can be formulated as a combinatorial optimization problem, where the objective is to identify an informative subset of features from a large search space (Guyon and Elisseeff, 2003; Liu and Yu, 2005). Therefore, MAS-based algorithms provide a natural framework for this task, as the collective behavior of agents enables efficient exploration of the feature space while balancing exploration and exploitation.

In the following subsection, we first describe how the feature selection problem can be represented as an appropriate search space for MAS. We then explain the different components of MAS and how they interact to explore this space effectively, highlighting how agent cooperation enables the discovery of informative feature subsets.

Representation of the feature selection problem

Let us assume a high-dimensional biological dataset $\mathbf{X} \in \mathbb{R}^{N \times D}$ with N patients, where each patient is represented by D biological features. Within the MAS framework, the search space for the feature selection problem can be represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which nodes correspond to features and edges represent possible transitions between features.

Beyond this structural representation, the graph incorporates two types of additional components that are essential for feature selection. First, the *static node score* and *static edge weight* serve as fixed measures that guide the selection of relevant features by quantifying their individual importance and pairwise associations. Specifically, the static node score, denoted as $\eta_{\mathcal{V}}(f_u)$, assigns a value to each node f_u based on the input data (e.g., feature relevance measures). Similarly, the static edge weight, denoted as $\eta_{\mathcal{E}}(f_u, f_v)$, assigns a value to each edge according to measures of correlation or dependency between features.

Second, the *dynamic node score* and *dynamic edge weight* evolve during the selection process, allowing agents to adaptively refine feature subsets over time. At iteration t, the dynamic node score of feature f_u is denoted by $\tau_u(t)$, and the dynamic edge weight between features f_u and f_v is denoted by $\tau_{uv}(t)$. These dynamic values are initialized with a constant c and iteratively updated as agents navigate the graph, reinforcing paths that contribute to promising feature subsets.

The framework of a general MAS algorithm

After representing the input data as a graph, MAS performs feature selection through an iterative improvement process. Each iteration consists of four key steps:

- **Step 1**: Initialize N_a agents, with each agent starting from a randomly selected node. The starting node corresponds to the first feature selected by that agent.
- **Step 2**: Each agent extends its current feature subset by performing a random walk over the search space. The movement follows *state transition rules*, where agents preferentially move toward nodes that are considered more desirable according to the model's criteria. This process continues until a predefined number of features has been selected.
- **Step 3**: Once all agents complete their walks, the resulting feature subsets are evaluated using a *fitness function*, such as classification performance or another chosen objective.
- **Step 4**: *Dynamic value updating rules* are applied to adjust the dynamic node scores and edge weights. Features and connections not selected experience decay, while those contributing to high-quality subsets receive reinforcement.

These steps are repeated for a maximum number of iterations. Finally, the best-performing feature subsets are selected. In the following, we describe each of these components in detail.

State transition rules

Each agent constructs a feature subset by exploring the feature space and selecting features based on either a greedy or a probabilistic transition rule. The *greedy rule* prioritizes the most promising feature at each step, while the *probabilistic rule* allows for stochastic exploration across different features.

We define the transition score from feature f_u to feature f_v at iteration t as a combination of static and dynamic node and edge values:

$$\textsc{Transition}(f_u, f_v) = \textsc{Combine}\Big(\eta_{\mathcal{V}}(f_v),\, \eta_{\mathcal{E}}(f_u, f_v),\, \tau_v(t),\, \tau_{uv}(t)\Big), \tag{2.23}$$

where $\textsc{Combine}(\cdot)$ is a model-specific function integrating these components. Under the greedy rule, agent a positioned at node f_u selects the next feature f_v such that

$$f_v = \underset{f_k \in \mathcal{J}_u(a)}{\arg\max} \Big[\textsc{Transition}(f_u, f_k)\Big], \quad \text{if } q \leq q_0, \tag{2.24}$$

where $\mathcal{J}_u(a)$ is the set of candidate features not yet selected by agent a, q is a random variable uniformly sampled from $[0,1]$, and q_0 is a predefined threshold.

Under the probabilistic rule, when $q > q_0$, the next feature f_v is selected according to the distribution

$$\text{Prob}(f_v \mid f_u) = \frac{\textsc{Transition}(f_u, f_v)}{\sum_{f_k \in \mathcal{J}_u(a)} \textsc{Transition}(f_u, f_k)}, \quad \text{if } q > q_0. \tag{2.25}$$

This mechanism balances *exploitation*, where agents deterministically select the most promising features, with *exploration*, where agents stochastically sample alternative features to diversify the search process (Dorigo and Stützle, 2018; Tabakhi et al., 2014).

Dynamic value updating rules

Dynamic value updating rules are applied to guide agent exploration and improve feature selection by modifying the dynamic node scores and edge weights based on their contributions

to high-quality feature subsets. For the next iteration $t+1$, these updates occur after all agents have constructed their feature subsets and are defined as:

$$\tau_u(t+1) = (1-\rho_{\mathcal{V}})\,\tau_u(t) + \rho_{\mathcal{V}}\,\textsc{Score}_{\mathcal{V}}(f_u), \tag{2.26}$$

$$\tau_{uv}(t+1) = (1-\rho_{\mathcal{E}})\,\tau_{uv}(t) + \rho_{\mathcal{E}}\,\textsc{Score}_{\mathcal{E}}(f_u, f_v), \tag{2.27}$$

where $\rho_{\mathcal{V}}, \rho_{\mathcal{E}} \in (0,1]$ are decay coefficients controlling the balance between retention of past information and incorporation of new evidence. The functions $\textsc{Score}_{\mathcal{V}}(\cdot)$ and $\textsc{Score}_{\mathcal{E}}(\cdot)$ compute update values for nodes and edges based on their relative contribution to the quality of the selected feature subsets. The specific definition of these functions depends on the model employed and is determined by task-specific evaluation criteria. This updating rule follows a reinforcement learning-inspired mechanism, where features and edges contributing to better feature subsets accumulate higher dynamic scores and weights, thereby improving their likelihood of selection in subsequent iterations (Dorigo and Gambardella, 1997; Dorigo and Blum, 2005).

2.2.5 Multimodal Learning

Many complex systems can only be understood by combining information from multiple perspectives of the same objects, or modalities, such as biological data, images, signals, text, or structured variables. *Multimodal learning* aims to develop predictive machine learning models that extract and integrate information from multiple data modalities (Baltrušaitis et al., 2018; Ektefaie et al., 2023). In this section, we formulate supervised multimodal omics learning and extend it to the case of missing modalities.

Definition 2.2.1. (Supervised Multimodal Omics Learning). In the context of multi-omics, supervised multimodal learning involves leveraging multiple biological data to build predictive models. Formally, let $\mathcal{D} = \{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^M\}$ represent the M omics modalities, where $\mathbf{X}^i \in \mathbb{R}^{N \times d_i}$ denotes the ith omics modality with N patients and d_i features. Given the corresponding label matrix $\mathbf{Y} \in \mathbb{R}^{N \times C}$ (e.g., cancer subtypes encoded as one-hot vectors), the

classification task is to learn a mapping $f_\theta : \mathcal{D} \rightarrow \mathbf{Y}$ that minimizes the loss function $\mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}})$, where $\hat{\mathbf{Y}} = f_\theta(\mathcal{D})$ represents the predicted labels. Here, θ is the set of learnable parameters of the predictive model f_θ, which are optimized during training to minimize the loss $\mathcal{L}$.

Definition 2.2.2. (Multiomics with Missing Modalities). In real-world multiomics datasets, some modalities may be missing for some patients, impacting the integration of multiple omics modalities. Formally, let $\mathbf{M} \in \{0,1\}^{N \times M}$ be a binary modality mask matrix, where $m_{ji} = 1$ indicates that the ith modality is available for the jth patient, and $m_{ji} = 0$ indicates its absence. For omics modality data i, $\mathbf{X}^i$, the jth patient's data is considered if $m_{ji} = 1$. The problem is to construct a model f that can robustly handle missing modalities in $\mathcal{D}$ while still learning the relationships across available modalities.

2.3 Overview of Multimodal Fusion Strategies for Complete Data

In the case where all modalities are available for all samples, subjects, or patients, multimodal fusion strategies focus on effectively combining complementary information across data sources to improve predictive performance. Figure 2.4 illustrates the categorization of multimodal learning models for multiomics data.

2.3.1 Feature-Level Fusion

Feature-level fusion aims to integrate extracted features from multiomics data to capture richer information. A learning model is then applied to the integrated features to perform downstream tasks (Singh et al., 2019; Ross, 2009). Generally, feature-level fusion comprises early fusion, transformation, factorization, and multimodal feature selection strategies, as described in the following sections.

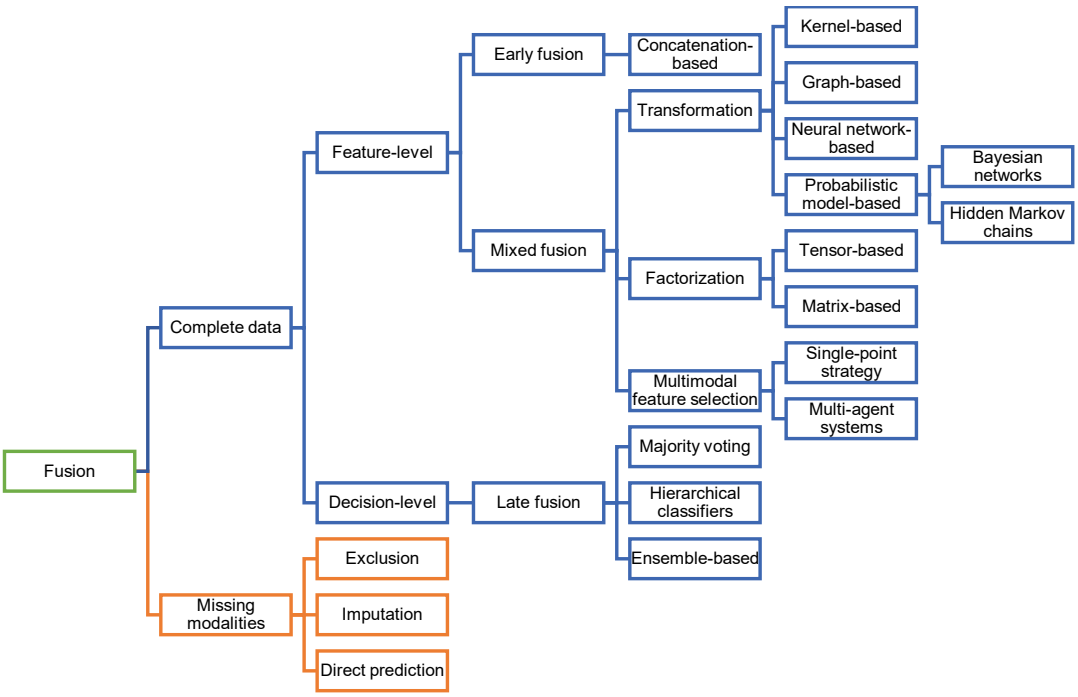

Fig. 2.4 **Categorization of multiomics data fusion approaches.**

Early fusion

The early fusion approach directly concatenates each omics dataset to construct a single comprehensive dataset containing all features before being fed into learning algorithms (Adossa et al., 2021). Methods based on this approach benefit from cross-modality learning, which refers to learning that involves information obtained from multiple modalities. This strategy is simple to develop, and the resulting joint dataset can be used as input for numerous classical machine learning algorithms, such as artificial neural networks (ANNs) (Kim et al., 2013), support vector machines (SVMs) (Ma et al., 2016), decision trees (DTs) (Lin and Lane, 2017), and random forests (RFs) (Ma et al., 2020). Training only one algorithm in early fusion leads to a more straightforward pipeline for implementing this approach (Baltrušaitis et al., 2018).

Despite its simplicity, the single dataset generated by early fusion often has a higher dimensionality relative to the number of patients. This problem makes the model's training difficult, decreases the performance, and increases the computational time (Ritchie et al.,

2015). Because there is a noticeable difference in feature dimensionality in multiomics data, another drawback of early fusion is the model's tendency to learn more from the omics modality with a larger number of features (Cavill et al., 2016). Employing dimensionality reduction techniques as a preprocessing step can mitigate these challenges by retaining a smaller set of discriminative features (Spicker et al., 2008; Wörheide et al., 2021; Mirza et al., 2019; Wang and Gu, 2016; Cantini et al., 2021).

Mixed fusion

In the mixed fusion approach, each omics dataset is first independently transformed into an intermediate representation, which is then merged to produce the final joint representation. Machine learning algorithms can subsequently be applied to this joint representation of multiomics data (Picard et al., 2021). Since the intermediate representations have a lower dimensionality, future analysis can be done more efficiently. Furthermore, using independent representations in the first step makes it possible to address the heterogeneity of multiple modalities (Ritchie et al., 2015). Transformation, factorization, and multimodal feature selection are three categories of mixed fusion, discussed in the following subsections.

Transformation. The goal of transformation-based methods is to project each unimodal dataset into a new subspace and combine these generated subspaces before building a learning model. These methods can integrate different omics modalities for diagnostic tasks (Lin and Lane, 2017). Strategies for generating the new subspace can be divided into four categories: kernel-based, graph-based, neural network-based, and probabilistic model-based fusions.

Kernel-based fusion is a well-known form of transformation that uses kernel functions to map original features onto a new space with higher dimensions. Kernels allow such methods to work in high-dimensional space to explore similarities and relationships between samples (Yan et al., 2017). Several widely used kernel functions include linear, polynomial, sigmoid, hyperbolic tangent, string, tree, graph, Gaussian, and radial basis functions (Roman et al., 2021). Kernel-based methods are capable of applying different kernels to multiple

omics datasets that provide various similarity metrics. SVM is a traditional machine learning algorithm for working with kernels (Cortes and Vapnik, 1995; Schölkopf et al., 2004; Ben-Hur et al., 2008). Multiple kernel learning (MKL) is another algorithm that utilizes different kernels for multiple omics modalities to find correlations across modalities and consolidates them into a single kernel for further analysis (Gönen and Alpaydın, 2011; Baltrušaitis et al., 2018; Lanckriet et al., 2004; Seoane et al., 2014). Other kernel-based fusion methods include semi-definite programming (SDP), SDP/SVM (Lanckriet et al., 2004), sequential minimal optimization MKL (SMO-MKL) (Tao et al., 2019), relevance vector machine (RVM) (Tipping, 2001), AdaBoost RVM (Wu et al., 2010), and kernel principal component analysis (PCA) (Mariette and Villa-Vialaneix, 2018).

Despite its attractive performance, this fusion approach is computationally expensive compared to other transformation-based techniques (Reel et al., 2021).

Graph-based fusion is becoming a prevalent technique for integrating multimodal data in biomedical and healthcare studies due to its capacity to capture molecular interactions (Li et al., 2022a). This type of fusion is broadly accomplished in two ways. The first way models each modality as a graph and combines them to create a unified graph for establishing further analysis (Chierici et al., 2020). In the graphs created for each omics, nodes represent samples, and edges represent relationships between pairs of samples. These graphs are subsequently converted to similarity matrices through an iterative optimization process (Wang et al., 2014) or a single iteration algorithm (Rappoport and Shamir, 2019). In the final stage, the matrices are fused to construct a single graph that can be fed into a machine learning model for performing learning tasks (Picard et al., 2021). Similarity network fusion (SNF) is a graph-based method that fuses constructed similarity graphs of patients through an iterative process of updating similarities to identify cancer subtypes derived from a clustering algorithm (Wang et al., 2014). Ranked SNF is an extension of SNF that uses a feature-ranking strategy for computing multiomics features' rankings to build a final graph before applying spectral clustering (Chierici et al., 2020). Rappoport and Shamir (2019) proposed a three-stage graph-based fusion algorithm for clustering called NEighborhood-based Multi-Omics (NEMO)

clustering. In the first stage, a similarity matrix was created for each omics modality based on patient relationships. Then, matrices were fused to generate a relative similarity matrix in a single iteration. Finally, the spectral clustering algorithm was used to cluster cancer samples. Ramirez et al. (2020), Kim et al. (2015), and Wen et al. (2021) have presented more fusion methods based on graph.

In contrast to the previous strategy that fuses unimodal graphs directly, graph representation learning integrates the latent representation spaces of graphs into a joint representation, which is then fed into machine learning models. In other words, each graph-structured input modality is encoded into a low-dimensional space that reflects the graph topology while preserving the original structure. The latent representations are then combined to perform the downstream task (Li et al., 2022a; Ektefaie et al., 2023). Amor et al. (2021) investigated multimodal learning on tissue-specific multiomics data using a graph embedding model based on the variational autoencoder. In the first phase, RNA sequencing and gene methylation datasets were independently transformed into compact vectorial spaces using two graph convolutional neural networks. These representations were then incorporated and fed into the variational graph autoencoder model for the purpose of link prediction. Zeng et al. (2019) proposed a graph-based fusion architecture using deep learning (deepDR) for in silico drug repositioning. They applied a random walk approach to convert each drug's structure into a vector representation, which was subsequently fused via a multimodal deep autoencoder for a prediction task. Graph information propagation network (GripNet) is a general framework to integrate several modalities using heterogeneous graph representation learning. In this framework, a new data structure named supergraph was defined to embed each modality in a compact space and pass messages between them for performing a specific task (Xu et al., 2023a).

Since graphs are formed based on samples rather than features, the complexity of the overall pipeline does not significantly increase when new omics modalities are added.

Neural network-based fusion is a growing approach in the multiomics research area due to its superior performance in numerous domains of multimodal learning (Ramachandram

and Taylor, 2017). In this approach, a network is trained on each modality from biological systems to learn a joint representation of the inputs. The hidden layers of the constructed networks are then passed into another neural network for further analysis (Mroueh et al., 2015). The benefits of using neural networks for fusion include hierarchical representations via layers of neural networks, the ability to learn complex non-linear relationships among features, and scalability with respect to the number of omics modalities (Kang et al., 2022). Bica et al. (2018) proposed a new neural network approach for the fusion of multiomics data derived from TCGA. They used two feed-forward neural networks, each receiving specific omics data, to obtain cross-correlations in multiple omics data and combine them into a fully connected network for the prediction tasks. Lee et al. (2019) and Alkhateeb et al. (2021) presented additional fusion methods based on neural networks.

An unsupervised class of neural network-based fusion methods is autoencoders, which learn compact representations of input omics data through an encoder–decoder structure. Poirion et al. (2018) introduced an algorithm in which an autoencoder is built for each omics data to link them for inferring survival subtypes. Zhang et al. (2018b) presented an unsupervised multiomics integration method based on an autoencoder with three hidden layers to identify prognosis subtypes.

Convolutional neural networks have been extended to graph-based fusion. The multi-omics graph convolutional network (MOGONET) is a multiomics fusion framework for classification that utilizes supervised convolutional networks for omics datasets (Wang et al., 2021b). The multi-omics graph convolutional network (M-GCN) is another multiomics fusion method developed based on convolutional neural networks for molecular subtyping (Yin et al., 2022).

Despite the strengths of neural network-based fusion methods, their performance depends on a large training sample size that is of limited availability in the multiomics field. Moreover, this approach lacks interpretability, an essential need for biologists seeking to identify biological functions (Bodein et al., 2022).

Probabilistic model-based fusion is commonly based on the hidden Markov model (HMM) to build probabilistic models that encode information as transition matrices and mix them for downstream tasks (Bayoudh et al., 2022). A Markov chain models variables (i.e., states) and the transition probabilities between states to produce a sequence of observations. The transition probabilities indicate the likelihood of moving from one state to another (Ghahramani and Jordan, 1995). The ability of HMMs to consider the correlations between states makes them effective for analyzing multiomics data (Yoon, 2009). The use of HMMs in biological analysis has been investigated in several works (Gentili et al., 2022; Yoon, 2009).

Another form of probabilistic model-based fusion is the Bayesian network approach, which constructs a directed acyclic graph to represent the probability distribution of each omics (Subramanian et al., 2020). Fridley et al. (2012) used a Bayesian hierarchical structure to fuse multiple types of genomics data with phenomics data to find the direct and indirect effects of genomics on the phenotype. Wang et al. (2019b) introduced a Bayesian framework to combine genomics, transcriptomics, and epigenomics data for identifying the high-confidence risk genes of schizophrenia. As another example, Zhang et al. (2022b) developed a machine learning method, called regional fine-mapping (RefMap), which is a hierarchical Bayesian framework for gene discovery in amyotrophic lateral sclerosis. In their work, epigenomics and transcriptomics have been integrated for gene discovery.

Factorization. Factorization-based fusion takes multiple omics modalities as input matrices and decomposes them into two components: (i) factors shared across all modalities and (ii) weights for each modality. Common factors can be utilized for patient clustering, and weights help identify biomarkers (Cantini et al., 2021). The decomposition assumes that biological mechanisms can be detected by biological factors shared among multiple modalities (Picard et al., 2021; Huang et al., 2017). Therefore, this type of fusion is capable of acquiring a complex inter-omics structure. Two approaches designed to perform factorization are discussed in the following subsections.

Matrix factorization fusion factorizes multiomics data matrices into the product of several matrices, including omics-specific weight matrices and a shared factor matrix (Cantini et al., 2021). As a result, data are projected into a shared latent space to identify driving factors of diseases. In unimodal learning, the most popular matrix factorization technique is PCA, which decomposes the covariance matrix of data to extract underlying biological factors (Bishop and Nasrabadi, 2006). Various methods have been developed to generalize PCA for multiomics fusion. Multi-omics factor analysis (MOFA) (Argelaguet et al., 2018), joint and individual variation explained (JIVE) (Lock et al., 2013), joint non-negative matrix factorization (jNMF) (Zhang et al., 2012), and integrative non-negative matrix factorization (iNMF) (Yang and Michailidis, 2016) are generalized PCA-based methods in which multiple omics modalities of the same biological samples are included in the analysis so features in each modality differ. In contrast, multi-study factor analysis (MSFA) (De Vito et al., 2019) is another generalization wherein the same omics features from different biological samples obtained in multiple studies are included in the analysis. Additionally, several other works have introduced integration methods based on matrix factorization, including iCluster (Shen et al., 2009) and its extension iCluster+ (Mo et al., 2013) by utilizing maximum likelihood estimation and regularized generalized canonical correlation analysis (RGCCA) (Tenenhaus and Tenenhaus, 2011).

Although matrix factorization has been extensively investigated in the literature, most existing methods assume a global shared latent space among omics modalities while neglecting partial common structures; a variable can be shared by two omics modalities but is not available in the third one (Yang and Michailidis, 2016; Gaynanova and Li, 2019). For example, Gaynanova and Li (2019) presented the structural learning and integrative decomposition (SLIDE) method to model partially shared structures for multi-view fusion.

Tensor-based factorization typically constructs higher-order relationships among biological variables to extract factors that play essential roles in describing these relationships (Liang et al., 2021). In other words, omics modalities are represented in a higher-dimensional space in which the new dimension indicates the data modality (Kolda and Bader, 2009).

Jung et al. (2021) introduced a two-stage tensor-based factorization method (MONTI) for multiomics analysis in cancer subtyping. In the first stage, non-negative tensor factorization was used to factorize tensors constructed from multiomics data, and in the second stage, a representative feature subset was selected using L_1 regularization.

Tensor-based factorization fusion can be computationally expensive, especially as the number of omics modalities increases (Kuleshov et al., 2015; Liang et al., 2021). Teschendorff et al. (2018) proposed a tensorial independent component analysis (tICA) based on independent component analysis. They used data tensors of order four to an epigenome-wide association study (EWAS) dataset, which resulted in better efficiency in comparison to the current methods. Moreover, a number of attempts have been made to use Bayesian inference in tensor factorization (Tang et al., 2018; Liu et al., 2022).

Multimodal feature selection. This approach selects features with joint consideration of multiple omics modalities during the integration process. In most existing multiomics fusion methods, feature selection is independently applied to each omics modality as a preprocessing step before integration. These methods reduce the feature space independently so that the relationships between multiple omics could be lost during such preprocessing (Picard et al., 2021).

Several works have explored multimodal feature selection for multiomics integration that can be categorized into single-point strategies and multi-agent systems.

The single-point strategy aims to select a subset of features from the whole omics data by starting from a specific point. This strategy iteratively adds new features selected from each omics modality according to a statistical metric. El-Manzalawy et al. (2018) presented a joint feature selection model applied to a multi-view cancer dataset. Their main idea was to generalize the minimum-redundancy and maximum-relevance statistical method, originally developed for single-view feature selection, to multiomics research through an incremental process.

Although this fusion strategy considers the correlations between omics modalities, its sensitivity to the starting feature is likely to result in limited performance.

Multi-agent systems are an improvement over the single-point strategy, in which several starting points are simultaneously selected to guide the feature selection procedure. In MAS, multiple independent agents collaborate within a shared environment to solve a complex problem. The distributed and parallel problem-solving abilities, using knowledge of other agents through interactions, decision-making flexibility of individual agents, and reliability are essential features of MAS to handle complex problems (Dorri et al., 2018).

2.3.2 Decision-Level Fusion

Fusion at the decision level (also known as late fusion) builds multiple machine learning models independently on each omics modality and then aggregates predictions from these models for the final decision. This approach is flexible because different machine learning models can be constructed for each omics modality (Adossa et al., 2021). Majority voting, hierarchical classifiers, and ensemble-based methods are the most extensively used aggregators at the level of decision (Carrillo-Perez et al., 2022; Sharifi-Noghabi et al., 2019; Huang et al., 2019; Miao et al., 2021).

The ability to integrate various single-omics frameworks to build multimodal learning algorithms is the key strength of decision-level fusion (Picard et al., 2021). In addition, having separate learning models enables this fusion approach to handle heterogeneity across multiple modalities. Because learning algorithms are independently trained on each omics data modality, late fusion can also address the feature imbalance problem.

2.4 Overview of Multimodal Fusion Strategies for Missing Modalities

Multimodal models for addressing missing modalities can be categorized into three approaches. The simplest approach excludes patients with missing modalities, applying methods designed for complete datasets (Wang et al., 2021b; Yang et al., 2022; Wang et al., 2014). While straightforward, this approach significantly reduces the number of patients and ignores

valuable information from those with missing modalities, particularly in healthcare where patients with complete modalities are limited (Pan et al., 2021; Wang et al., 2020).

Imputation offers an alternative by filling or reconstructing missing modalities using various strategies. One strategy concatenates all modalities, treating missing modalities as missing data at random and estimating them based on existing values (Rubin, 1996; Schafer, 1999; Austin et al., 2021). However, this strategy ignores the natural correlation of features within the same modality (Mitra et al., 2023; Zhang et al., 2022a). Another strategy uses statistical methods that account for the structure of missing modalities. TOBMI (Dong et al., 2019) applies a weighted kNN method to impute block-wise missing multiomics. M3Care (Zhang et al., 2022a) imputes missing modalities in the latent space by identifying similar patients and concatenates the imputed representations to perform clinical tasks. Deep generative models provide another imputation strategy by reconstructing missing modalities from available ones. Common models include autoencoders (Ngiam et al., 2011; Gong et al., 2021; Ashuach et al., 2023), generative adversarial networks (Cai et al., 2018; Shang et al., 2017), and diffusion models (He et al., 2024). Another example is SMIL (Ma et al., 2021), which utilizes Bayesian meta-learning to reconstruct missing modalities. These models may fail to distinguish genuinely missing modalities from zeros (Collier et al., 2020), assume equal importance for all modalities (Yao et al., 2024), introduce imputation noise due to a limited number of patients (Zhang et al., 2022a; Wang et al., 2020), or rely on strong data distribution assumptions (Wu et al., 2024b).

Direct prediction is another approach that incorporates specialized designs to perform downstream tasks with missing modalities. NEMO (Rappoport and Shamir, 2019) manages missing modalities using a neighborhood-based approach with average similarities. MRGCN (Yang et al., 2023) uses a GCN-based encoder-decoder method with an indicator matrix to address missing modalities. GRAPE (You et al., 2020) and MUSE (Wu et al., 2024b) leverage bipartite graphs with modalities and patients as nodes to directly address missing modalities. DrFuse (Yao et al., 2024) combines modality-specific and shared sub-models. MT (Ma et al., 2022a) addresses cases where a specific modality is entirely missing in test data. ASR (Chen et al., 2022) is a sparse representation-based method that learns modality-specific sparse

representations from available data and fuses them into a unified similarity matrix in an unsupervised setting. However, these methods may oversimplify relationships, struggle with multimodal missingness scalability due to the exponentially growing number of missing patterns, or address only specific missing scenarios (Zhang et al., 2022a; Wu et al., 2024b).

Chapter 3

Multimodal Feature Selection Using Multi-Agent Systems

Building upon the foundational concepts of graphs and multi-agent systems introduced in Chapter 2, this chapter addresses **Research Question 1**, as outlined in Section 1.2. Here, we focus on developing a multimodal feature selection approach to reduce omics feature dimensionality while accounting for both intra- and inter-omics relationships. To this end, we propose `MAgentOmics`, a **M**ultimodal feature selection framework based on multi-**Agent** systems (MAS) for multi**Omics** integration. To the best of our knowledge, this is the first approach to employ a multi-agent system for feature selection in a multiomics dataset. We evaluate the proposed method on a publicly available dataset from TCGA and demonstrate its effectiveness in the cancer classification task.

3.1 Introduction

The advent of high-throughput technologies has recently generated massive amounts of biological omics data, enabling comprehensive assessment of molecular systems and creating new opportunities for biologists and data scientists to diagnose, treat, and even cure cancers (Barefoot et al., 2019; Tong et al., 2020). Multiomics data integration is key to cancer prediction, as it captures different aspects of molecular mechanisms (Hassanzadeh et al.,

2015; Krassowski et al., 2020). However, biological data usually suffer from the small sample size problem because data collection is financially costly and the number of clinical research participants is limited (Acosta et al., 2022; Picard et al., 2021). Therefore, most omics datasets contain a large number of features compared to only a relatively small number of patients, leading to the classical phenomenon in machine learning known as the *curse of dimensionality* (De Meulder et al., 2018; Cantini et al., 2021). Although many features are available in omics data, their correlations are very high, and some features do not highlight disease-specific indicators (Hira and Gillies, 2015; Kirpich et al., 2018). It becomes even more challenging when irrelevant and redundant features in each omics modality are integrated into the multiomics analysis. Therefore, these features mislead the learning algorithm and limit the model's generalizability to unseen samples. Moreover, the computational complexity of developing a model is substantially increased in the presence of high-dimensional datasets.

Since multiomics data are highly dimensional, many studies have utilized feature selection to simplify the integration process (Chen et al., 2020; Wang et al., 2021b; Zhang et al., 2019b; Torres-Martos et al., 2022; Lualdi and Fasano, 2019; Goh and Wong, 2016; Smit et al., 2008). However, most existing feature selection methods are independently applied to each omics dataset as a preprocessing step, thereby neglecting cross-omics relationships. As discussed in Section 2.3.1, some efforts have been made to develop multimodal feature selection methods (El-Manzalawy et al., 2018), but these methods typically identify feature subsets in a single-iteration process that starts from a specific point. Consequently, they are prone to becoming trapped in a local optimum.

We propose `MAgentOmics`, an MAS framework for multimodal feature selection that jointly considers multiple omics modalities. To the best of our knowledge, this is the first approach to employ an MAS for multimodal feature selection to address the high dimensionality inherent in multiomics data. In this framework, each omics modality is represented as a feature-level graph, where nodes correspond to features and edges capture their relationships. The graphs from all modalities are then connected to form a unified search space, referred to as the *multiomics feature network*, which incorporates both intra- and

inter-omics interactions. `MAgentOmics` comprises agents that interact by sharing knowledge during the search process. Iteratively, each agent generates a subset of features based on relevance and redundancy analyses, and a new fitness function is introduced to evaluate the constructed solutions. To integrate information across modalities, a probability distribution is defined to represent the relative importance of each modality, which is dynamically updated through a new *omics importance updating rule*. Finally, the global-best solution is selected as the final feature subset. `MAgentOmics` performs feature selection in an unsupervised manner, without relying on patient labels.

We evaluate the performance of `MAgentOmics` on publicly available multiomics data from TCGA. The results demonstrate that `MAgentOmics` outperforms an existing multimodal feature selection method, highlighting the advantages of integrating multiple omics modalities within an MAS framework.

3.2 Contributions

Our main contributions can be summarized as follows:

- We develop a multimodal feature selection method based on an MAS that operates on graph representations of feature spaces across all omics modalities. This work extends the conventional MAS algorithm from single-modality datasets to multimodal settings through the introduction of new components.
- We design `MAgentOmics` to mitigate the curse of dimensionality by modeling both intra- and inter-omics interactions. This is achieved through the introduction of the multiomics feature network, a graph-based framework.
- We introduce a new fitness function and an omics importance updating rule specifically designed for multimodal feature selection.

3.3 Methodology

`MAgentOmics` is an MAS-based architecture for feature selection, designed for multiomics data. We first describe how the search space is modeled as a suitable graph for MAS. Then, we explain the different components of the MAS for multimodal feature selection and how they interact to explore this space effectively.

3.3.1 Multiomics Feature Network Representation

Feature selection in multiomics data requires a structured representation that captures both feature relationships within each omics modality and interactions across different modalities. To achieve this, we extend the representation introduced in Section 2.2.4 by proposing the *multiomics feature network*, a graph-based framework in which features from different omics modalities are modeled as nodes and their relationships as edges. This formulation enables multi-agent exploration for feature selection. Figure 3.1 illustrates the proposed network representation.

Formally, the multiomics feature network represents multiomics data as a complete weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where $\mathcal{V} = \mathcal{V}^1 \cup \cdots \cup \mathcal{V}^M = \bigcup_{i=1}^{M} \mathcal{V}^i$ is the set of nodes representing features from all omics modalities, with each modality i contributing a feature set $\mathcal{V}^i = \{f_1^i, \ldots, f_{d_i}^i\}$. $\mathcal{E}$ is the set of edges connecting related features, capturing both intra-omics and cross-omics relationships. Moreover, $\mathcal{W} = \{w_1, \ldots, w_M\}$ denotes the relative importance of omics modalities, where w_i represents the importance of omics modality i.

Beyond the structural representation, this graph incorporates two key components essential for feature selection. First, the *static node score* and *static edge weight* serve as fixed measures that guide the selection of relevant features by quantifying their individual significance and pairwise connections. Specifically, static node score, denoted as $\eta_{\mathcal{V}}(f_u^i)$, assigns a score to each node f_u^i, computed using the variance (Theodoridis and Koutroumbas, 2008). Variance measures the variability of a feature across patients, with higher variability indicating more discriminative information. Similarly, static edge weight, denoted as $\eta_{\mathcal{E}}(f_u^i, f_v^i)$, assigns a value to each edge based on the absolute value of the Pearson correlation

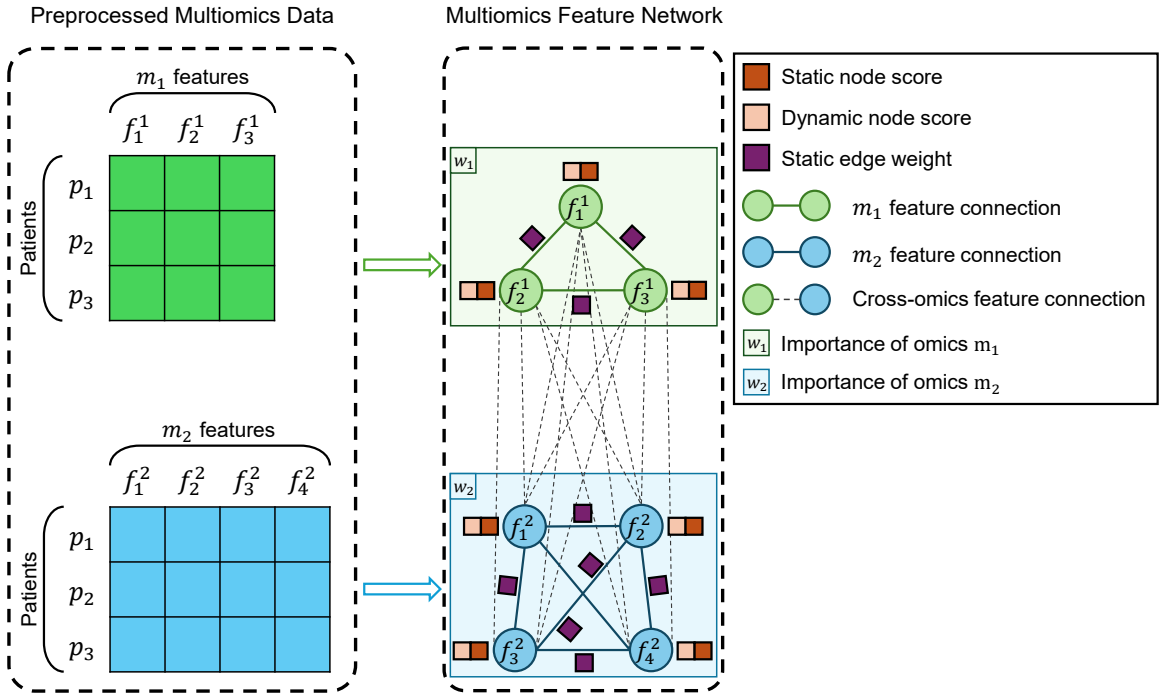

Fig. 3.1 **Network-based representation of multiomics data for feature selection.** Each omics modality is modeled as a network, where nodes represent omics-specific features and edges connect feature pairs. Cross-omics connections link each feature in one omics modality to all features in the other. Static node score quantifies feature importance based on variance, while static edge weight represents the Pearson correlation between feature pairs. The dynamic node score evolves based on agent interactions throughout the feature selection process.

(Theodoridis and Koutroumbas, 2008). Second, the *dynamic node score*, which evolves during feature selection, guides agents in exploring and refining feature subsets over time, thereby increasing the attractiveness of good solutions as the process progresses. $\tau_u^i(t)$ denotes the dynamic node score of feature node f_u^i at time t. The dynamic value starts from a constant $c_\mathcal{V}$ and is iteratively adjusted as agents navigate the graph.

3.3.2 The `MAgentOmics` Framework

Initially, the static node score of each feature is computed separately using variance. Then, the static edge weights between pairs of features within each omics modality are calculated and assigned to the corresponding graph edges. Thereafter, the initial dynamic node score is

set to a small constant c. Finally, the initial probability values are equally distributed across omics modalities ($w_i = 1/M, \forall i = 1,2,\ldots,M$).

The feature selection procedure in `MAgentOmics` follows an iterative process, where each iteration of the algorithm consists of five steps.

- **Step 1**: For each modality i, N_a agents are placed on randomly selected graph nodes within that modality as their initial positions. The starting node represents the first feature selected by each agent.

- **Step 2**: Each agent constructs a candidate feature subset by performing a random walk over the multiomics feature network $\mathcal{G}$. The movement follows *state transition rules*, where agents preferentially move toward nodes connected by low-correlation edges and having high static and dynamic node scores. These rules enable agents to greedily search for subsequent features within their current omics modality or probabilistically explore other modalities for potential features. This process continues until a predefined number of features, N_f, are selected.

- **Step 3**: Once all agents complete their walks, the resulting feature subsets are evaluated using a new *fitness function*. The subset with the highest fitness score is then retained as the current-best solution.

- **Step 4**: The *dynamic value updating rule* is applied to modify the dynamic node scores. Unselected features experience decay, while those contributing to the selected feature subsets receive increased dynamic values. In other words, features that are repeatedly selected by agents are more likely to be chosen in subsequent iterations and assigned higher dynamic values. Moreover, the agent holding the current-best solution can allocate an additional dynamic value to its selected features.

- **Step 5**: The *omics importance updating rule* is applied to dynamically modify the selection probability of each omics modality in subsequent iterations, increasing the likelihood of exploring those that contain informative features. This rule is inspired by

reinforcement learning and encourages agents to explore states that have yielded high positive reinforcement over time (Mitchell, 1997).

These steps are repeated for a maximum of T iterations. The global-best solution is selected from the best solutions obtained across all iterations. Each omics dataset is then reduced to the features represented in the global-best solution, and the final dataset is constructed by combining the reduced omics datasets. In the following sections, we describe each of these components in detail.

3.3.3 State Transition Rules

Each agent constructs a feature subset by exploring the multiomics feature network and selecting features based on either a greedy or probabilistic transition rule. The greedy rule prioritizes the most promising feature at each step, while the probabilistic rule allows for stochastic exploration across different omics modalities. Figure 3.2 illustrates the available feature choices under each rule with possible transitions, as well as an example of a complete constructed solution.

We define the transition score from feature f_u^i to feature f_v^i at iteration t as a combination of static and dynamic node and edge values:

$$\textsc{Transition}(f_u^i, f_v^i) = \frac{\tau_v^i(t)\,\eta_{\mathcal{V}}(f_v^i)}{\eta_{\mathcal{E}}(f_u^i, f_v^i)}. \tag{3.1}$$

The core intuition is to guide agents to walk through the multiomics feature network by favoring features that are individually informative through the static node score and supported by their past contributions via the dynamic node score. Transitions between highly correlated features are discouraged through the static edge weight. As a result, the transition mechanism promotes the exploration of features that are both informative and complementary to those already selected, leading to more diverse and effective feature subsets.

When following the greedy strategy, agent a, positioned at node f_u^i, selects the next feature f_v^i within the same omics modality i at iteration t based on the highest transition

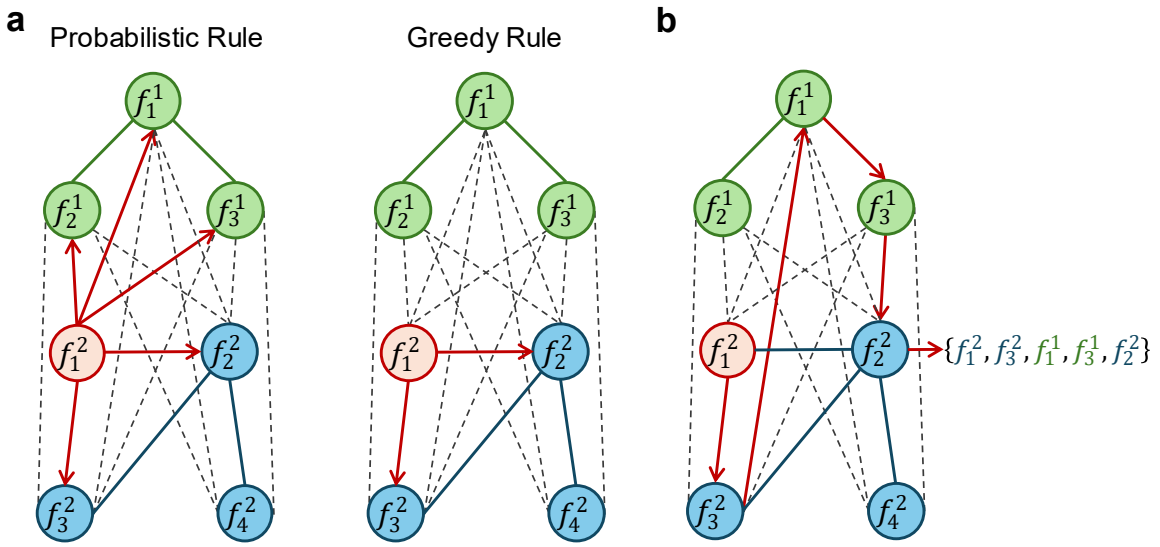

Fig. 3.2 **An example of state transition rules. a**, An agent placed at node f_1^2 can move probabilistically or greedily to another node, following one of the possible edges shown by red arrows. **b**, The agent constructs its solution iteratively using state transition rules, forming a candidate feature subset starting from node f_1^2.

score:

$$f_v^i = \underset{f_k^i \in \mathcal{J}_{u^i}^i(a)}{\arg\max} \left[\textsc{Transition}(f_u^i, f_k^i)\right], \quad \text{if } q \leq q_0, \tag{3.2}$$

where $\mathcal{J}_{u^i}^i(a)$ is the set of candidate features that agent a has not yet selected within omics modality i, q is a random variable uniformly sampled from $[0,1]$, and q_0 is a predefined threshold.

In the probabilistic rule, the agent stochastically explores the network by first selecting an omics modality and then transitioning to a new feature within that modality. Given an agent positioned in modality i, it randomly chooses another modality j from the set of omics modalities $\mathcal{M}$ (where j can be equal to i) based on the omics importance weights $\mathcal{W}$, as follows:

$$j = \textsc{Choice}(\mathcal{M}, \mathcal{W}), \quad \text{if } q > q_0. \tag{3.3}$$

Once the new modality j is selected, the next feature v is determined by:

$$\text{Prob}(f_v^j \mid f_u^i) = \frac{\textsc{Transition}(f_u^i, f_v^j)}{\sum_{f_k^j \in \mathcal{J}_{u^i}^j(a)} \textsc{Transition}(f_u^i, f_k^j)}. \tag{3.4}$$

3.3.4 Dynamic Value Updating Rule

We apply this rule to guide agent exploration and improve feature selection by modifying the dynamic node scores according to their contributions to high-quality feature subsets. This update is performed after all agents have constructed their feature subsets and is computed as follows:

$$\tau_u^i(t+1) = (1-\rho_{\mathcal{V}})\,\tau_u^i(t) + \rho_{\mathcal{V}} \left[\frac{\text{Count}_{\mathcal{V}}(\{f_u^i\})}{\text{Count}_{\mathcal{V}}(\mathcal{S}_{\mathcal{V}})} + \left[\Delta\tau_u^i(t)\right]^{\text{best}} \right], \tag{3.5}$$

where

$$\left[\Delta\tau_u^i(t)\right]^{\text{best}} = \begin{cases} \textsc{Fitness}(\mathcal{S}(\text{best})), & \text{if } f_u^i \in \mathcal{S}_{\mathcal{V}}(\text{best}), \\ 0, & \text{otherwise}, \end{cases} \tag{3.6}$$

$\rho_{\mathcal{V}} \in (0,1]$ is a decay coefficient that controls the balance between retaining past information and incorporating new evidence. The function $\text{Count}_{\mathcal{V}}(\cdot)$ determines how frequently a node has been selected by agents during the current iteration, and $\mathcal{S}_{\mathcal{V}}$ denotes the set of all selected features in the current iteration. In Equation (3.6), $\textsc{Fitness}(\mathcal{S}(\text{best}))$ represents the fitness function used to evaluate the quality of the subset $\mathcal{S}(\text{best})$, which corresponds to the current-best solution.

The dynamic value update acts as the collective memory of the multi-agent system. This update rewards features that are frequently selected by agents and contribute to strong candidate subsets, while providing a bonus to features that belong to the current-best subset. The decay term ensures that features which lose their relevance over time are gradually selected less frequently. This balances exploration and exploitation during the search process, allowing agents to adapt to more promising regions of the feature space as iterations progress.

3.3.5 Fitness Function

The fitness function evaluates the quality of solutions. In this study, we introduce a fitness function that incorporates two key metrics commonly used in feature selection, maximal relevance and minimal redundancy. To assess the quality of the solution $\mathcal{S}(a)$ constructed by agent a, we propose the following quantitative measure:

$$\textsc{Fitness}\big(\mathcal{S}(a)\big) = \frac{\text{Mean}_{f_u^i \in \mathcal{S}(a)}[\eta_{\mathcal{V}}(f_u^i)]}{\text{Mean}_{\{f_u^i, f_v^j\} \subset \mathcal{S}(a)}[\eta_{\mathcal{E}}(f_u^i, f_v^j)]}, \tag{3.7}$$

where the numerator computes the average static node score of the features in the selected subset $\mathcal{S}(a)$ (i.e., the average relevance calculated using variance), and the denominator computes the average static edge weight between each pair of distinct features in $\mathcal{S}(a)$ (i.e., the average redundancy calculated using Pearson correlation).

The design of this fitness function is rooted in the principles of maximum relevance and minimum redundancy in feature selection (Peng et al., 2005; Liu and Yu, 2005). The rationale is to define a quality score that guides the search toward better feature subsets by maximizing the discriminative information of individual features while simultaneously minimizing redundancy between them. By placing relevance in the numerator and redundancy in the denominator, we construct a metric that assesses the quality of candidate feature subsets. The numerator promotes the selection of features with high individual importance, while the denominator penalizes subsets containing highly correlated features. As a result, the fitness value increases when the selected features are both informative and minimally redundant. Conversely, subsets with low relevance or high redundancy yield lower fitness values. This formulation encourages agents to identify compact feature subsets that maximize information content while avoiding redundant features.

3.3.6 Omics Importance Updating Rule

As we define the relative importance of omics modalities, w_i, we introduce a function to update these importance values over time. This rule increases the weight of omics modalities

that contain informative features for agents in subsequent iterations. The following equation is proposed and implemented to update the omics importance:

$$w_i(t+1) = (1-\gamma)\, w_i(t) + \gamma \left[\frac{\text{Count}_{\mathcal{V}}(\mathcal{S}_{\mathcal{V}}^i)}{\text{Count}_{\mathcal{V}}(\mathcal{S}_{\mathcal{V}})} \right], \tag{3.8}$$

where $\gamma \in (0,1]$ controls the update rate of the omics importance values, $w_i(t)$ and $w_i(t+1)$ denote the importance values of omics i at time t and $t+1$, respectively, and $\mathcal{S}_{\mathcal{V}}^i$ represents the subset of selected features from omics i during the current iteration.

The intuition behind this design is to provide a simple yet effective mechanism that enables the framework to automatically identify and prioritize informative modalities without relying on prior biological assumptions. The update rule follows a weighted update scheme, where the importance of each modality is adjusted based on its contribution to the selected feature subsets over time. Modalities that consistently contribute more features to the selected subsets gradually receive higher weights, while those contributing fewer features experience a reduction in their importance. The parameter γ controls how quickly the weights adapt, balancing the influence of past importance values and current observations.

3.3.7 The `MAgentOmics` Algorithm

Algorithm 3.1 outlines the framework of `MAgentOmics`. This framework comprises three main sections, including initialization (lines 1–6), the multimodal feature selection procedure (lines 7–18), and final dataset construction (lines 19–20).

3.4 Experiments

3.4.1 Data Collection

The performance of `MAgentOmics` is evaluated on a publicly available dataset from the TCGA program (Weinstein et al., 2013). We select ovarian serous cystadenocarcinoma (OV) data from TCGA and analyze three omics modalities, including gene-level copy number variation (CNV), DNA methylation (DNA), and gene expression RNAseq (mRNA). The OV

Algorithm 3.1 `MAgentOmics`: Multimodal Feature Selection Using Multi-Agent Systems

Input

$\mathcal{D} =< (\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^M), \mathbf{Y} >$: multiomics dataset with M omics modality
T: predefined number of iterations
N_A: the number of agents per omics modality

Output

$\mathcal{D}' =< (\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^M), \mathbf{Y} >$: Reduced-dimension multiomics dataset

for $i = 1$ to M **do**
    Compute static node score $\eta_{\mathcal{V}}(f_u^i)$
    Compute static edge weight $\eta_{\mathcal{E}}(f_u^i, f_v^i)$
    Initialize dynamic node score $\tau_u^i(0) \leftarrow c_{\mathcal{V}}$
    Initialize omics importance $w_i \leftarrow \frac{1}{M}$
end for
for $t = 1$ to T **do**
    for $i = 1$ to M **do**
        for $a = 1$ to N_A **do**
            Place agent a randomly on a unique node
            Build a solution by iteratively applying Equations (3.2), (3.3), and (3.4)
            Evaluate the solution with the fitness function in Equation (3.7)
        end for
    end for
    Retain the solution with the highest fitness as $\mathcal{S}(\text{best})$ in iteration t
    Update dynamic node scores using Equation (3.5)
    Apply omics importance updating rule using Equation (3.8)
end for
Choose the global-best solution found
Build the multiomics dataset $\mathcal{D}'$ based on the global-best solution

dataset is downloaded from the UCSC Xena Platform.[1] A brief description of the dataset is provided in Table 3.1.

For the classification task, the data are divided into two groups based on clinical information (i.e., recorded time of death and vital status) in the TCGA ovarian cancer dataset, namely, the *long-term* and *short-term* groups. The long-term group includes patients with a survival time ≥ 3 years. On the other hand, the short-term group includes patients with a survival time < 3 years and vital status of "DECEASED" (El-Manzalawy et al., 2018). Since data from all omics modalities are not measured for each patient, only those patients with all omics modalities measured and complete clinical information are included in this

[1] `https://xenabrowser.net/datapages/`.

study. Therefore, the final multiomics dataset contains 176 patients, of whom 85 belong to the long-term group.

Table 3.1 Characteristics of ovarian multiomics data used in the `MAgentOmics` experiments.

Omics Modality	Version	#Features	#Patients
DNA	2017-09-08	27,578	616
CNV	2017-09-08	24,776	579
mRNA	2017-10-13	20,530	308

3.4.2 Data Preprocessing

The raw OV data are preprocessed by the following three steps to ensure robustness in the classification task (Zhang et al., 2021; Vangimalla and Sreevalsan-Nair, 2021). First, features with missing values in any omics modality are removed to ensure a consistent and complete feature space, particularly given the high dimensionality and limited sample size of the OV dataset (El-Manzalawy et al., 2018). Second, feature values are normalized to the range $[0, 1]$ using min-max normalization (You et al., 2020), ensuring comparability across features from different modalities with varying scales (Schulte-Sasse et al., 2021; Wu et al., 2024a). This also ensures that the static node scores and edge weights are comparable across modalities. Finally, features with variance lower than 0.05 are filtered out, as they show limited variation across samples and are unlikely to contribute to distinguishing patterns. This threshold is chosen to eliminate near-constant features while retaining sufficient variability for downstream learning. A higher threshold risks removing informative features and reducing diversity, thereby limiting the discovery of complementary information, whereas a lower threshold unnecessarily expands the computational search space for the agents.

3.4.3 Baseline

We empirically evaluate the performance of `MAgentOmics` as an unsupervised feature selection method in comparison with mRMR-mv (El-Manzalawy et al., 2018), a supervised multimodal feature selection method. mRMR-mv extends the minimum-redundancy maximum-

relevance (mRMR) (Peng et al., 2005) statistical method, originally developed for feature selection for a single modality, to multiomics research through an incremental process.

3.4.4 Evaluation Metric

Since OV is a well-balanced dataset, we use accuracy to evaluate and compare the performance of different methods. In our experiments, the accuracy obtained from two widely used classifiers, logistic regression (LR) (Le Cessie and Van Houwelingen, 1992) and random forest (RF) (Breiman, 2001), is used as the performance metric.

3.4.5 Evaluation Strategy

To evaluate the average classification accuracy of the selected subsets, we use 5-fold cross-validation (CV). Since k-fold CV can produce a noisy estimate of classifier performance, we repeat the CV procedure five times and report the average accuracy across all runs.

3.4.6 Implementation Details

`MAgentOmics` is implemented in Python 3.10, incorporating essential functionalities from pandas 1.3.4 (The pandas development team, 2021) and NumPy 1.20.3 (Harris et al., 2020). We use scikit-learn 0.24.2 (Pedregosa et al., 2011) to implement the LR and RF classifiers. For `MAgentOmics`, the following parameters are used: the maximum number of iterations $T = 30$, the number of agents per omics modality $N_A = 20$, the decay coefficients in Equations (3.5) and (3.8), $\rho_{\mathcal{V}} = 0.2$ and $\gamma = 0.2$, the initial dynamic node score for each node $\tau_u^i(0) = 0.2$, and the state transition rule control parameter $q_0 = 0.7$. The code for `MAgentOmics` is publicly available on GitHub.[2]

For the mRMR-mv baseline method, the absolute value of Pearson's correlation coefficient is used as the redundancy function. Moreover, mutual information implemented in scikit-learn is used as the relevance function, following the original paper (El-Manzalawy et al., 2018).

[2] `https://github.com/SinaTabakhi/MAgentOmics`.

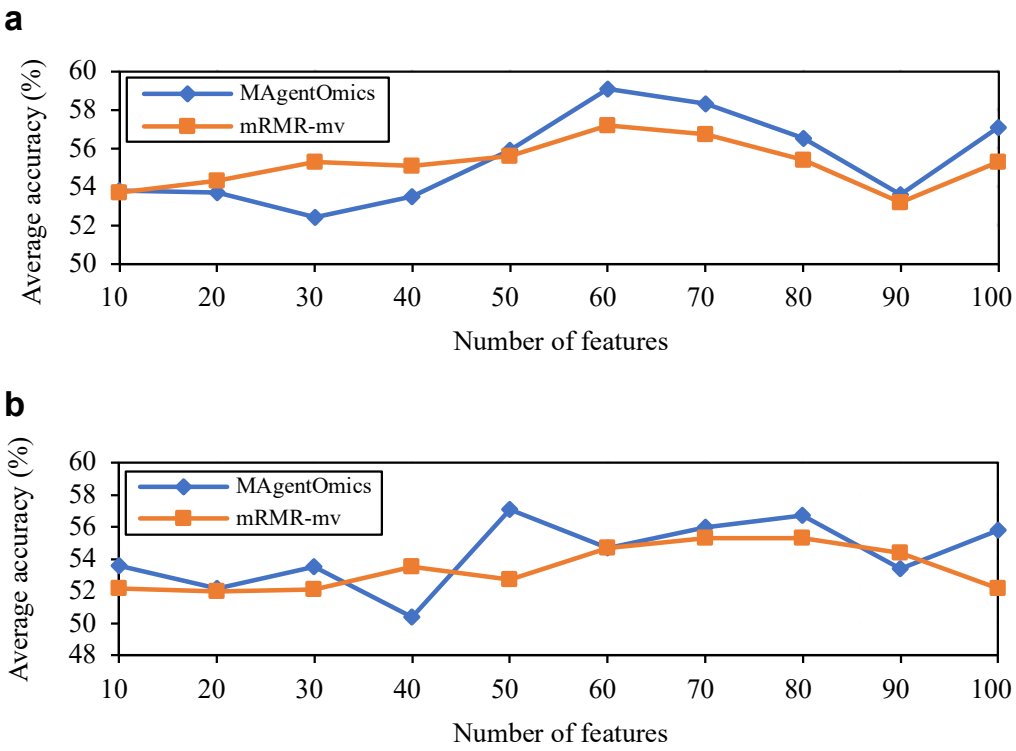

Fig. 3.3 **Classification performance comparison (average accuracy) across different sizes of selected feature subsets using 5-fold cross-validation. a**, Results for logistic regression. **b**, Results for random forest.

3.4.7 Classification Performance Comparison

Figure 3.3 compares the classification performance of `MAgentOmics` and mRMR-mv in terms of average accuracy (in %) obtained using logistic regression (LR) and random forest (RF) classifiers across different numbers of selected features. As shown in Figure 3.3a, `MAgentOmics` outperforms mRMR-mv when the number of selected features is 10, 50, 60, 70, 80, 90, and 100. Similarly, Figure 3.3b shows that the classification accuracy of `MAgentOmics` is generally higher than that of mRMR-mv. For instance, when 50 features are selected, `MAgentOmics` achieves an accuracy of 57.1%, compared to 52.7% for mRMR-mv.

Overall, the best performance of `MAgentOmics` is 59.1% and 57.1% for LR and RF, respectively, while mRMR-mv achieves 57.2% and 55.3% for the same classifiers.

We limit the number of selected features from 10 to 100 to balance predictive performance and model complexity. While a larger number of features can often be helpful for biological

interpretation, increasing the number of features in MAgentOmics introduces two primary challenges: (1) an increased risk of overfitting due to the high dimensionality relative to the number of samples (Ruiz et al., 2023; Cantini et al., 2021), and (2) higher computational costs as the search space expands significantly. Moreover, feature selection aims to identify compact and informative subsets, and including a larger number of features may reduce the impact of the most relevant features (Acosta et al., 2022; El-Manzalawy et al., 2018). While this chapter focuses on the methodological efficiency of MAgentOmics, a more extensive analysis of modality-specific contributions and larger feature sets is provided in Chapter 4.

3.5 Summary

In this chapter, we proposed MAgentOmics, a multimodal feature selection method based on a multi-agent system designed for multiomics data integration. The architecture relies on interactions among multiple agents that share their knowledge to identify an optimal global feature subset. Moreover, we introduced a new fitness function to evaluate the generated subsets based on the principles of maximum relevance and minimum redundancy. To assess the relative importance of each modality in the feature selection process, we developed an omics importance updating rule inspired by reinforcement learning.

The performance of MAgentOmics was compared with the supervised mRMR-mv feature selection method in terms of classification accuracy using LR and RF classifiers. MAgentOmics achieved promising results on the TCGA ovarian cancer dataset compared with mRMR-mv. Although the proposed fitness function evaluates the relevance and redundancy of candidate feature subsets, it requires high computational time to calculate redundancies. Furthermore, as it is an unsupervised function, it may yield lower accuracy for some datasets.

Future work will focus on designing new fitness functions that can evaluate subsets more accurately with lower computational cost. Since the number of edges between pairs of features is large in multiomics data, employing a sparse graph representation of the search space could improve execution time.

Chapter 4

Multimodal Learning with Heterogeneous Graphs

In Chapter 3, we introduced a multimodal feature selection method to reduce omics feature dimensionality while accounting for both intra- and inter-omics relationships. In this chapter, we leverage that approach, which provides auxiliary and essential structural information for constructing heterogeneous graphs, addressing **Research Question 2** as outlined in Section 1.2. To this end, we propose `HeteroGATomics`, a **Hetero**geneous **G**raph **AT**tention network for **omics** integration aimed at improving cancer diagnosis. We first enhance our `MAgentOmics` algorithm by generating sparse graphs in the feature space rather than fully connected ones and by introducing dynamic edge weights as additional information. This feature similarity network in the feature space is then combined with a patient similarity network in the patient space. The resulting combination forms a dual-view framework that automatically constructs modality-specific heterogeneous graphs. We validate the performance of `HeteroGATomics` on three cancer datasets and explore its interpretability through biomarker identification.

4.1 Introduction

Building upon concepts introduced in earlier chapters, recent advancements in sequencing technologies have significantly accelerated the generation of multimodal biological data, collectively known as multiomics. This progress enables personalized medicine by constructing comprehensive patient profiles across multiple omics modalities (Steyaert et al., 2023; Acosta et al., 2022). While each omics modality contains valuable information, their collective integration enables new insights into the fundamental aspects of human disease biology, particularly in cancer research (Karczewski and Snyder, 2018; Wang et al., 2021b). Research initiatives developing integrative multiomics models (Cantini et al., 2021; Acosta et al., 2022; Schulte-Sasse et al., 2021) leverage the unique and complementary characteristics of each modality to construct multimodal models that are more accurate and interpretable than unimodal models, thereby enhancing biological predictions and facilitating biomarker discovery (Karczewski and Snyder, 2018; Li et al., 2022a; Schulte-Sasse et al., 2021).

As pointed out in Section 2.2.2, GNNs (Hamilton et al., 2018) have been increasingly utilized for multiomics integration (Li et al., 2022a; Chen et al., 2025; Forster et al., 2022; Ektefaie et al., 2023; Wang et al., 2021b). GNNs offer more accurate models with improved decision-making power by effectively capturing both the intra- and inter-omics structures within the data. Modeling each omics modality as a graph, where biological entities are nodes and their interactions are edges, facilitates a deeper understanding of entity interactions (Li et al., 2022a).

GNNs in multiomics analysis, while promising, still face two key challenges. The first challenge is learning from high-dimensional data, where each omics suffers from having a large number of features compared to a small patient cohort (Steyaert et al., 2023). This issue hinders the ability of GNNs to learn meaningful representations, reducing performance in biomedical applications (Cantini et al., 2021). Most current studies apply feature selection strategies independently to each omics modality (Picard et al., 2021; Wang et al., 2021b), without accounting for relationships among them, which can compromise interpretability and model performance. A few studies have explored multimodal feature selection across all omics (El-Manzalawy et al., 2018), yet these often involve either a single-iteration greedy

process or creating fully connected graphs rather than sparse graphs in feature space, leading to reduced predictive capability or high computational demands. Therefore, there is a clear gap in the development of better methods that effectively account for intra- and inter-omics relationships in multiomics.

The second challenge is the limited expressive power of conventional GNNs developed for homogeneous graphs. In these models (Zheng et al., 2024; Wu et al., 2024a; Wang et al., 2021b; Forster et al., 2022; Schulte-Sasse et al., 2021), the learning process incorporates only patient or feature similarity networks with a single type of node and edge. Representing multiple input modalities as a homogeneous graph loses crucial structural information inherent in the data, fails to account for the diverse nature of the data, and limits the model's ability to capture complex biological interactions. In contrast, heterogeneous graphs offer a solution to these limitations by modeling multiple types of nodes and edges to capture the diversity of the data. While some methods have extended learning to heterogeneous graphs (Zitnik et al., 2024; Ruiz et al., 2023; Xu et al., 2023a), they often rely on pre-existing knowledge graphs to represent semantic relationships between different entities, such as genes, patients, and diseases. However, such pre-existing knowledge graphs may not always be available, and creating them often requires domain knowledge or expertise and can be costly (Chandak et al., 2023).

In this chapter, we present `HeteroGATomics`, a novel framework employing heterogeneous GAT for integrating multiomics data for cancer diagnosis. `HeteroGATomics` operates in two distinct stages: multimodal feature selection and heterogeneous graph learning. Unlike previous feature selection methods, `HeteroGATomics` implements an MAS for a multimodal feature selection on sparse graphs, constructed across the feature spaces of all omics. This approach creates a *feature similarity network* for each modality, representing one view of the input data. The proposed multimodal feature selection strategy not only achieves competitive performance but also captures structural features arising from MAS. Moreover, for each omics modality, a *patient similarity network* is built, forming the second data view. We propose a dual-view approach to automatically construct modality-specific heterogeneous graphs by connecting feature and patient similarity networks, followed by representation

learning with GAT encoders to generate predictions. This heterogeneous graph construction improves the expressive power of GATs by capturing diverse structural information. Leveraging late fusion for final decision-making in `HeteroGATomics` integrates predictions from all modalities in a supervised manner for multiomics integration (Wang et al., 2021b).

We conduct comprehensive experiments to evaluate `HeteroGATomics`' diagnosis performance on three cancer multiomics datasets. The results show that `HeteroGATomics` consistently outperforms baseline methods, highlighting the benefits of integrating diverse omics data. Additionally, we show the necessity of each module of `HeteroGATomics` via a series of ablation studies. We further explore `HeteroGATomics`' interpretability through a biomarker identification process, revealing its ability to identify cancer-related biomarkers. By pinpointing interacting biomarker networks and highlighting key cancer-related functions, it also identifies potential therapeutic targets.

4.2 Contributions

Our main contributions can be summarized as follows:

- We enhance our previous `MAgentOmics` algorithm for multimodal feature selection to operate on sparse graphs. We also introduce a dynamic edge weight value to the multiomics feature network introduced in Section 3.3.1. The state transition rules are updated accordingly, and a new fitness function is proposed to directly incorporate the performance of the selected subsets evaluated by a classifier. Moreover, we present a dynamic value updating rule for updating the dynamic edge weight.
- We propose a dual-view approach to construct heterogeneous graphs and learn holistic graph representations that capture both patient- and feature-level structures.
- We identify cancer-related biomarkers enabled by interpretable multimodal fusion.

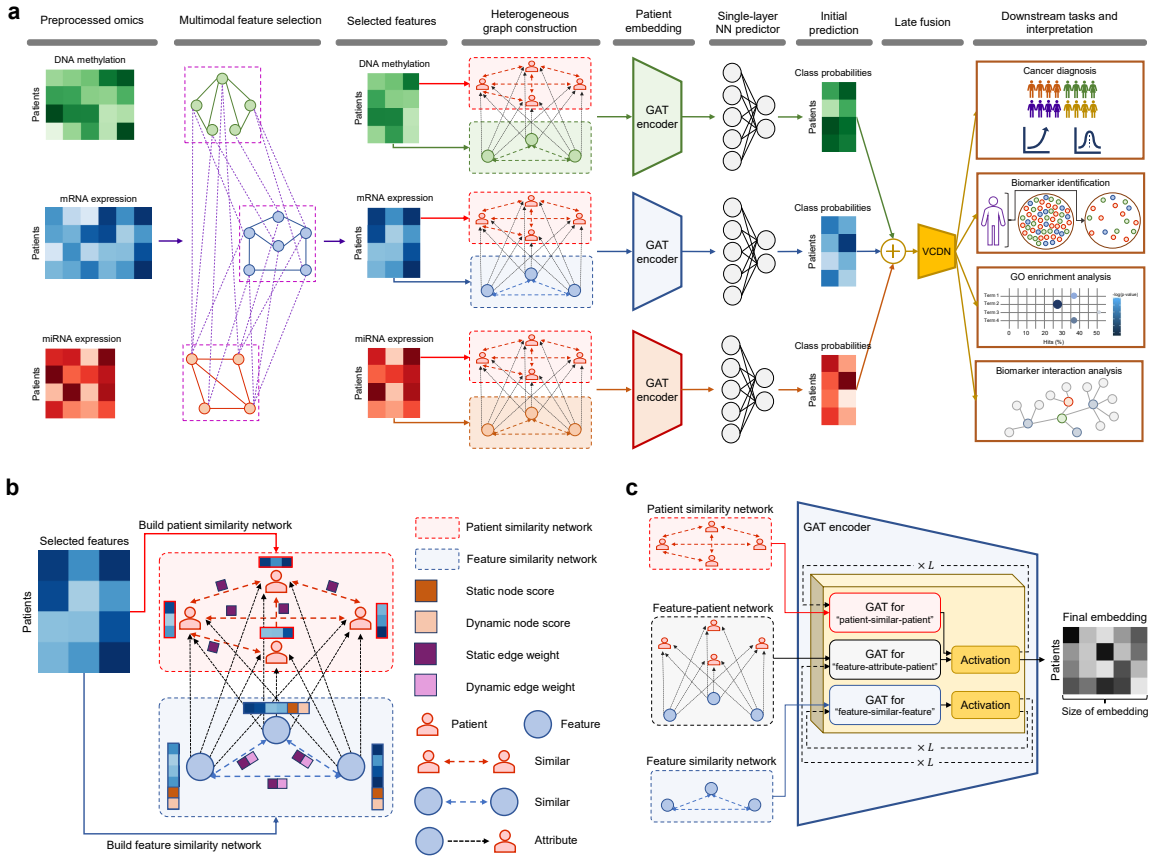

Fig. 4.1 `HeteroGATomics` **architecture. a**, `HeteroGATomics` integrates multimodal feature selection and heterogeneous graph learning in six steps. (1) `HeteroGATomics` represents the preprocessed omics as feature similarity networks, where each network represents a specific omics with nodes corresponding to features and edges denoting their correlations. All omics modalities are interconnected at the raw feature level to capture cross-modality interactions. (2) An MAS performs multimodal feature selection on these networks to select informative features, considering both intra- and cross-modality interactions. (3) `HeteroGATomics` builds a patient similarity network for each omics and combines it with the feature similarity network to construct a heterogeneous graph. (4) GAT encoders learn the representations of each individual heterogeneous graph. (5) A single-layer neural network predicts patient labels from the learned representations. (6) A late fusion combines predicted labels from all modalities and feeds them into a VCDN network to perform downstream tasks. **b**, The heterogeneous graph construction combines feature and patient similarity networks through feature-patient relations. **c**, Multiple stacked GAT layers (denoted by L) encode the heterogeneous graph into hidden representations for each node type. Each layer uses three GATs to learn the three relations within the graph, updating node representations by aggregating relation-specific information.

4.3 Methodology

In this section, we formally introduce the overall framework of `HeteroGATomics`, as shown in Figure 4.1a. `HeteroGATomics` performs supervised multiomics integration with two main modules: an MAS for dimensionality reduction of preprocessed omics and a GAT architecture for heterogeneous graph representation learning.

`HeteroGATomics` first represents each preprocessed omics as a sparse feature similarity network, where each node corresponds to an individual feature and each edge indicates the correlation between a pair of features. Then, it uses the MAS algorithm to select features in a multimodal manner from all omics modalities, leveraging both intra- and cross-modality interactions at the feature level to utilize complementary information in multiomics datasets.

After multimodal feature selection, `HeteroGATomics` creates a patient similarity network for each omics, where nodes represent patients and edges represent correlations between their features. Next, it constructs a heterogeneous graph for each omics modality by connecting the patient and feature similarity networks, connecting feature nodes to all patient nodes (Figure 4.1b). This helps capture patient-level and feature-level relationships in a unified representation, providing a comprehensive view of the dataset. Then, a heterogeneous GAT model encodes the structures inherent in each input heterogeneous graph and learns meaningful node representations (Figure 4.1c). Afterward, a single-layer fully connected neural network takes the learned node representations as input to predict cancer. Finally, a late fusion strategy consolidates predictions across modalities by aggregating the generated predictions and feeding them into a view correlation discovery network (VCDN) (Wang et al., 2021b) for final prediction. The model architecture and implementation details are presented in the following subsections.

To the best of our knowledge, `HeteroGATomics` is the first method to explore cross-modality interactions at both the feature level, using the MAS algorithm, and the label level, employing a late fusion strategy. Moreover, the multimodal feature selection algorithm in `HeteroGATomics` not only reduces feature dimensionality but also provides more structural information for the heterogeneous graph, benefiting downstream tasks.

4.3.1 Multimodal Feature Selection

The high dimensionality of multiomics datasets results in many irrelevant and redundant features, making it challenging for GNNs to learn meaningful representations and perform accurate classification. Therefore, an effective feature selection method is required to identify representative features. We propose a multimodal feature selection strategy, an improved version of our `MAgentOmics` algorithm, to mitigate the curse of dimensionality by modeling both intra- and cross-omics interactions, instead of applying separate feature selection to each omics modality.

Multiomics feature network representation

The input omics modalities are represented as a multiomics feature network, as described in Section 3.3.1. The multiomics feature network, $\mathcal{G}$, integrates feature similarity networks $\mathcal{G}^i$ constructed for each omics modality i. Due to the high dimensionality and the presence of correlated features within each $\mathcal{G}^i$, employing a sparse graph rather than a fully connected one in the feature selection algorithm accelerates training, reduces memory usage, and improves computational efficiency. Therefore, we retain only a percentage of edges whose weights fall below a threshold θ_f. Furthermore, the static node score of each feature, $\eta_{\mathcal{V}}(f_u^i)$, is evaluated using ANOVA and assigned to the corresponding node, making `HeteroGATomics` a supervised feature selection method, unlike `MAgentOmics`. This shift to supervised feature selection better aligns feature selection with the classification task. While `MAgentOmics` provides a general framework independent of labels, `HeteroGATomics` focuses on classification tasks, where discriminative features with respect to class labels are essential, which cannot be guaranteed by unsupervised criteria alone. By using ANOVA to incorporate label information, we provide a more informative feature space for the GAT to learn complex interactions necessary for downstream prediction.

In addition to these components, we introduce a dynamic value $\tau_{uv}^i(t)$, representing the edge weight between feature nodes f_u^i and f_v^i at time t. This value is initialized with a constant $c_{\mathcal{E}}$ and iteratively updated as agents navigate the graph.

Algorithm 4.1 `HeteroGATomics`: Multimodal Feature Selection Module

Input

$\mathcal{D} = <(\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^M), \mathbf{Y}>$: multiomics dataset with M omics modality
T: predefined number of iterations
N_A: the number of agents per omics modality

Output

$\mathcal{D}' = <(\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^M), \mathbf{Y}>$: reduced-dimension multiomics dataset
$\{\tau_u^i, \tau_{uv}^i\}_{i=1}^M$: dynamic node scores and edge weights

for $i = 1$ to M **do**
    Compute static node score $\eta_{\mathcal{V}}(f_u^i)$
    Compute static edge weight $\eta_{\mathcal{E}}(f_u^i, f_v^i)$
    Initialize dynamic node score $\tau_u^i(0) \leftarrow c_{\mathcal{V}}$
    Initialize dynamic edge weight $\tau_{uv}^i(0) \leftarrow c_{\mathcal{E}}$
    Initialize omics importance $w_i \leftarrow \frac{1}{M}$
    Build a sparse graph by retaining edges with weights below θ_f
end for
for $t = 1$ to T **do**
    for $i = 1$ to M **do**
        for $a = 1$ to N_A **do**
            Place agent a randomly on a unique node
            Build a solution by iteratively applying Equations (4.3), (3.3), and (4.4)
            Evaluate the solution with the fitness function in Equation (4.7)
        end for
    end for
    Retain the solution with the highest fitness as $\mathcal{S}(\text{best})$ in iteration t
    Update dynamic node scores using Equation (3.5)
    Update dynamic edge weights using Equation (4.5)
    Apply omics importance updating rule using Equation (3.8)
end for
Select the top B features based on weighted static and dynamic node scores
Build the multiomics dataset $\mathcal{D}'$ with the top B features

MAS for multimodal feature selection in `HeteroGATomics`

`HeteroGATomics` follows the overall procedure for performing feature selection as described in the `MAgentOmics` framework (Section 3.3.2) with modifications to accommodate its architecture. Algorithm 4.1 presents the framework of `HeteroGATomics` for multimodal feature selection. The following sections describe the components that differ from the `MAgentOmics` framework.

State transition rules. The state transition rules determine how the next feature is selected to extend the current feature subset solution, either greedily or probabilistically at each construction step. We define the transition score from feature f_u^i to feature f_v^i at iteration t as follows:

$$\textsc{Transition}(\mathcal{S}_{\mathcal{V}}(a), f_v^i) = \tau_v^i(t) + \eta_{\mathcal{V}}(f_v^i) - \textsc{Similarity}(\mathcal{S}_{\mathcal{V}}(a), f_v^i), \tag{4.1}$$

where

$$\textsc{Similarity}(\mathcal{S}_{\mathcal{V}}(a), f_v^i) = \text{Mean}_{f_u^i \in \mathcal{S}_{\mathcal{V}}(a)} \big[\eta_{\mathcal{E}}(f_u^i, f_v^i) \big], \tag{4.2}$$

$\mathcal{S}_{\mathcal{V}}(a)$ denotes the set of features currently selected by agent a. In Equation (4.2), the function $\textsc{Similarity}(\mathcal{S}_{\mathcal{V}}(a), f_v^i)$ measures the average correlation between the candidate feature f_v^i and the selected features.

When following the greedy strategy, agent a, positioned at node f_u^i, selects the next feature f_v^i within the same omics modality i at iteration t based on the highest transition score:

$$f_v^i = \underset{f_k^i \in \mathcal{J}_{u^i}^i(a)}{\arg\max} \big[\textsc{Transition}(\mathcal{S}_{\mathcal{V}}(a), f_k^i) + \tau_{uk}^i(t) \big], \quad \text{if } q \leq q_0. \tag{4.3}$$

In the probabilistic rule, an agent positioned in modality i first randomly selects another modality j according to Equation (3.3), then chooses the next feature v in that modality based on

$$\text{Prob}(f_v^j \mid f_u^i) = \frac{\textsc{Transition}(\mathcal{S}_{\mathcal{V}}(a), f_v^j)}{\sum_{f_k^j \in \mathcal{J}_{u^i}^j(a)} \textsc{Transition}(\mathcal{S}_{\mathcal{V}}(a), f_k^j)}. \tag{4.4}$$

In Equation (4.4), excluding the dynamic edge weight values simplifies the model and reduces computational cost.

Dynamic value updating rules. Dynamic node scores are updated as described in Equation (3.5). With the introduction of the new component $\tau_{uv}^i(t)$ as a dynamic edge weight, these values are updated after all agents have constructed their feature subsets as follows:

$$\tau_{uv}^{i}(t+1) = (1-\rho_{\mathcal{E}})\,\tau_{uv}^{i}(t) + \rho_{\mathcal{E}}\left[\frac{\mathrm{Count}_{\mathcal{E}}\big(\{(f_u^i, f_v^i)\}\big)}{\mathrm{Count}_{\mathcal{E}}(\mathcal{S}_{\mathcal{E}})} + \big[\Delta\tau_{uv}^{i}(t)\big]^{\text{best}}\right], \tag{4.5}$$

where

$$\big[\Delta\tau_{uv}^{i}(t)\big]^{\text{best}} = \begin{cases} \textsc{Fitness}\big(\mathcal{S}(\text{best})\big), & \text{if } (f_u^i, f_v^i) \in \mathcal{S}_{\mathcal{E}}(\text{best}), \\ 0, & \text{otherwise}, \end{cases} \tag{4.6}$$

$\rho_{\mathcal{E}} \in (0,1]$ is a decay coefficient that controls the balance between retaining past information and incorporating new evidence. The function $\mathrm{Count}_{\mathcal{E}}(\cdot)$ measures how frequently an edge has been selected by agents during the current iteration, and $\mathcal{S}_{\mathcal{E}}$ denotes the set of all selected edges in that iteration.

Fitness function. To evaluate the quality of the solution $\mathcal{S}(a)$ constructed by agent a, we define the following quantitative measure:

$$\textsc{Fitness}(\mathcal{S}(a)) = \frac{1}{3}\Big[\textsc{Quality}(\mathcal{S}(a)) + \textsc{Penalty}(\mathcal{S}(a))\Big], \tag{4.7}$$

where

$$\textsc{Penalty}(\mathcal{S}(a)) = \mathrm{Mean}_{f_u^i \in \mathcal{S}_{\mathcal{V}}(a)}\big[\eta_{\mathcal{V}}(f_u^i)\big] - \mathrm{Mean}_{(f_u^i, f_v^i) \in \mathcal{S}_{\mathcal{E}}(a)}\big[\eta_{\mathcal{E}}(f_u^i, f_v^i)\big], \tag{4.8}$$

$\textsc{Quality}(\cdot)$ quantifies the classifier's performance on the subset of selected features $\mathcal{S}(a)$. The term $\mathcal{S}_{\mathcal{V}}(a)$ denotes the set of selected features, while $\mathcal{S}_{\mathcal{E}}(a)$ represents the set of edges within the agent's solution. The function $\textsc{Penalty}(\cdot)$ serves as a regularization term that penalizes irrelevant and redundant features, encouraging the discovery of more informative and representative subsets.

4.3.2 Heterogeneous Graph Learning

After feature selection reduces feature dimensionality, we construct a heterogeneous graph for each omics. This graph is automatically generated, incorporating extensive structural information inherent in tabular omics for downstream tasks. Following this, a GAT encodes

each heterogeneous graph, effectively representing node information. The powerful attention mechanism in the GAT architecture prioritizes important nodes and edges, which is beneficial for omics data where specific genes influence biological processes or disease mechanisms. This focus on key elements in GAT enhances performance compared to other GNN architectures (Veličković et al., 2018; Forster et al., 2022; Hamilton, 2020). After encoding each heterogeneous graph, a single-layer neural network performs label prediction. We combine predictions from multiple omics into a tensor representing cross-omics label correlations, which is then processed through a VCDN for final prediction. Algorithm 4.2 presents the pseudo-code for the proposed architecture.

Heterogeneous graph construction

For each omics, we construct a heterogeneous graph by combining a feature similarity network with a patient similarity network (see Figure 4.1b). The feature similarity network, deriving from the feature selection phase, consists of nodes representing selected features and edges denoting the correlations between them. Node attributes are defined by their static and dynamic node scores, along with the corresponding values from the input omics, while edge weights are determined by the static and dynamic edge weights between nodes. Complementing this, we construct the patient similarity network where nodes represent individual patients, and edges denote the dynamic edge weight between them. These correlations, quantified using the absolute Pearson correlation coefficient, serve as edge weights. Only edges with weights above a specified threshold, θ_s, are maintained, which filters for the most significant patient correlations and results in a more manageable and relevant graph structure.

These two networks, indicating two distinct node types (patients and features), are interconnected to form a comprehensive heterogeneous graph encompassing three specific relations: *patient–similar–patient*, *feature–similar–feature*, and *feature–attribute–patient*. Each relation in the heterogeneous graph offers unique insights: *patient–similar–patient* reveals shared disease characteristics among patients; *feature–similar–feature* highlights the correlated dynamics of features in biological processes; and *feature–attribute–patient* provides key understanding of how individual features impact patient outcomes. This

Algorithm 4.2 `HeteroGATomics`: Heterogeneous Graph Construction and Learning Module

Input

$\mathcal{D}' = <(\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^M), \mathbf{Y}>$: reduced-dimension multiomics dataset
$\{\tau_u^i, \tau_{uv}^i\}_{i=1}^M$: dynamic node scores and edge weights
T_{pre}: number of epochs for pre-training
T_{train}: number of epochs for training

Output

$\hat{\mathbf{Y}} \in \mathbb{R}^{N \times C}$: final predicted labels

for $i = 1$ to M **do**
    Build patient similarity network from $\mathbf{Z}^i$
    Build feature similarity network using τ_u^i, τ_{uv}^i, and $\mathbf{Z}^i$
    Build heterogeneous graph $\mathcal{G}^i$ from feature and patient similarity networks
end for
for $t = 1$ to T_{pre} **do**
    for $i = 1$ to M **do**
        Apply GAT to learn representations on graph $\mathcal{G}^i$ via Equation (4.14)
        Predict omics-specific labels using a single-layer neural network
        Optimize omics-specific model parameters via Equation (4.15)
    end for
end for
for $t = 1$ to T_{train} **do**
    for $i = 1$ to M **do**
        Apply GAT to learn representations on graph $\mathcal{G}^i$ via Equation (4.14)
        Predict omics-specific labels using a single-layer neural network
    end for
    Create a cross-omics discovery tensor for each patient via Equation (4.17)
    Predict final labels $\hat{\mathbf{Y}}$ with VCDN from the created tensor
    Optimize VCDN parameters by minimizing the cross-entropy loss according to Equation (4.18)
    for $i = 1$ to M **do**
        Optimize omics-specific model parameters via Equation (4.15)
    end for
end for

approach of relation-aware representation in the graph allows for capturing more detailed information, reflecting the diverse characteristics of the target nodes in omics datasets.

Formally, for each omics modality i, we define a heterogeneous graph $\mathcal{G}^i = (\mathcal{V}^i, \mathcal{E}^i, \phi, \psi)$ as described in Section 2.2.1.

Representation learning and label prediction

`HeteroGATomics` employs multiple GAT layers to encode omics-specific heterogeneous graphs, where each GAT within a layer is tailored to a specific relation type (as illustrated in Figure 4.1c). For a given heterogeneous graph $\mathcal{G}^i$, the goal of representation learning is to develop three relation-specific functions $f_1 : \mathcal{V}_1^i \to \mathbb{R}^{d_h}$, $f_2 : \mathcal{V}_1^i \to \mathbb{R}^{d_h}$, and $f_3 : \mathcal{V}_2^i \to \mathbb{R}^{d_h}$, each mapping the nodes involved in a particular relation into a d_h-dimensional embedding space. The node types are partitioned into two disjoint types, $\mathcal{V}^i = \mathcal{V}_1^i \cup \mathcal{V}_2^i$, where $\mathcal{V}_1^i \cap \mathcal{V}_2^i = \emptyset$. When a node is the destination of several relations, their corresponding representations are aggregated. The `HeteroGATomics` encoder stacks multiple layers of three individual GATs, each encoding source nodes for a specific relation. For a node u, layer $l+1$ aggregates messages from its relation-typed neighborhoods as follows:

$$\mathbf{m}_{r,u\leftarrow v}^{(l)} = \alpha_{r,uv}^{(l)} \mathbf{W}_r^{(l)} \mathbf{h}_v^{(l)}, \tag{4.9}$$

$$\mathbf{m}_{\mathcal{N}(u)}^{(l)} = \frac{1}{|\mathcal{R}_u|} \sum_{r\in\mathcal{R}_u} \sum_{v\in\mathcal{N}_r(u)} \mathbf{m}_{r,u\leftarrow v}^{(l)}, \tag{4.10}$$

$$\mathbf{h}_u^{(l+1)} = \sigma\left(\mathbf{m}_{\mathcal{N}(u)}^{(l)}\right), \tag{4.11}$$

where $\mathcal{R}_u$ is the set of all relations for which node u serves as the destination, $\mathcal{N}_r(u)$ denotes the neighbors of node u under relation r, $\sigma(\cdot)$ represents the LeakyReLU nonlinearity, and $\mathbf{W}_r^{(l)}$ is the relation-specific learnable weight matrix. Self-loops are added to $\mathcal{N}_r(u)$ for both the *patient–similar–patient* and *feature–similar–feature* relations. In Equation (4.9), $\alpha_{r,uv}^{(l)}$ denotes the normalized attention coefficient between a pair of nodes u and v under relation r, which is calculated as follows:

$$\alpha_{r,uv}^{(l)} = \frac{\exp(e_{r,uv}^{(l)})}{\sum_{k\in\mathcal{N}_r(u)} \exp(e_{r,uk}^{(l)})}, \tag{4.12}$$

where

$$e_{r,uv}^{(l)} = \sigma\left(\mathbf{a}_{r,d}^{(l)^\top} \mathbf{W}_{r,d}^{(l)} \mathbf{h}_u^{(l)} + \mathbf{a}_{r,s}^{(l)^\top} \mathbf{W}_{r,s}^{(l)} \mathbf{h}_v^{(l)} + \mathbf{a}_{r,e}^{(l)^\top} \mathbf{W}_{r,e}^{(l)} \mathbf{e}_{uv}\right), \tag{4.13}$$

$\mathbf{e}_{r,uv}$ indicates the edge attributes between nodes u and v under relation r, and $\mathbf{a}_{r,s}^{(l)}$, $\mathbf{a}_{r,d}^{(l)}$, and $\mathbf{a}_{r,e}^{(l)}$ are the trainable attention vectors to weigh source, destination, and edge attributes, correspondingly. In Equation (4.12), $\exp(\cdot)$ denotes the standard exponential function and in Equation (4.13), $\sigma(\cdot)$ represents the LEAKYRELU nonlinearity.

To stabilize the attention mechanism learning process, multi-head attention (Veličković et al., 2018) is widely utilized. This scheme combines the outputs of K independent attention mechanisms to form the final node representations, defined as:

$$\mathbf{h}_u^{(l+1)} = \frac{1}{K}\sum_{k=1}^{K}\sigma\big(\mathbf{m}_{\mathcal{N}(u)}^{(l,k)}\big), \tag{4.14}$$

where $\mathbf{m}_{\mathcal{N}(u)}^{(l,k)}$ denotes the aggregated message from the neighborhood of node u computed by the k-th attention head, each with its own parameters.

Once each heterogeneous graph is mapped into node embeddings through attention layers, a single-layer neural network uses them for omics-specific label prediction. We optimize all trainable parameters via backpropagation to minimize the cross-entropy loss across all training patients, defined as:

$$\mathcal{L}_i = \sum_{(\mathbf{x}_u^i,\mathbf{y}_u)\in\mathcal{D}_{\text{tr}}^i} \text{CE}_i(\mathbf{y}_u, \hat{\mathbf{y}}_u^i), \tag{4.15}$$

where

$$\text{CE}_i(\mathbf{y}_u, \hat{\mathbf{y}}_u^i) = -\sum_{c=1}^{C} y_{uc}\log\frac{\exp(\hat{y}_{uc}^i)}{\sum_{k=1}^{C}\exp(\hat{y}_{uk}^i)}. \tag{4.16}$$

$\mathcal{D}_{\text{tr}}^i$ represents all the training patients in omics i, $\mathbf{x}_u^i$ is the uth training patient with its corresponding label $\mathbf{y}_u$, and $\hat{\mathbf{y}}_u^i$ is the class-probability vector predicted by the GAT model. The function $\text{CE}_i(\cdot,\cdot)$ represents the cross-entropy loss. In Equation (4.16), C is the number of classes, and $\hat{y}_{uc}^i$ represents the cth element in the predicted probability vector.

We utilize VCDN, which integrates omics-specific label predictions to perform the final classification. This network effectively learns both intra- and cross-modality correlations within the label space, thereby enhancing performance across several tasks (Wang et al., 2019a, 2021b). To utilize this network, we construct a cross-modality discovery tensor, T_u,

for each patient u by integrating the predicted class probabilities from all modalities. Each element of T_u is computed as follows:

$$T_{u,t_1t_2\cdots t_M} = \prod_{i=1}^{M} \hat{y}^i_{ut_i}, \qquad t_i \in \{1,2,\ldots,C\}, \tag{4.17}$$

where $\hat{y}^i_{ut_i}$ denotes the t_ith element of the predicted class-probability vector $\hat{\mathbf{y}}^i_u$ from modality i, and M is the total number of modalities. The resulting tensor T_u is reshaped into a vector of dimension C^M. VCDN, implemented as a fully connected network, takes this vector as input and outputs class probabilities for the final prediction. The network is trained using the cross-entropy loss function defined as:

$$\mathcal{L}_{\text{VCDN}} = \sum_{(\mathbf{y}_u, T_u)\in\mathcal{D}_{\text{tr}}} \text{CE}_{\text{VCDN}}(\mathbf{y}_u, \text{VCDN}(T_u)), \tag{4.18}$$

where $\text{VCDN}(T_u)$ is the network output for the tensor T_u of patient u in the training set $\mathcal{D}_{\text{tr}}$.

Computational complexity

The heterogeneous graph for each omics modality consists of patient and feature node types. In multiomics datasets, the number of patients, N, is relatively small, while each omics modality i has a high-dimensional feature space, denoted by d_i. The feature selection module reduces this dimensionality by retaining a subset of informative features across all modalities. For each modality, a subset of $B_i \ll d_i$ informative features is selected, resulting in a heterogeneous graph with $N + B_i$ nodes and making GAT learning more manageable. The construction of the patient similarity network involves pairwise comparisons between patients and remains computationally feasible due to the limited number of patients, despite its $\mathcal{O}(N^2)$ complexity, and is performed as a one-time preprocessing step. Since the number of edges also significantly affects computational cost, sparsification is applied to both feature and patient similarity networks to retain informative connections while reducing the overall computational burden.

4.4 Experiments

In this section, we evaluate the performance of `HeteroGATomics` across multiple datasets, perform ablation studies to assess the impact of individual modules, and provide a biological interpretation of the results.

4.4.1 Data Collection

We evaluate `HeteroGATomics` performance across three multiomics datasets derived from TCGA cohort, for two downstream tasks, including bladder urothelial carcinoma (BLCA) grade classification, brain lower grade glioma (LGG) grade classification, and renal cell carcinoma (RCC) subtype classification. We download all datasets from the TCGA cohort via the UCSC Xena platform (Goldman et al., 2020), which provides DNA methylation (DNA), gene expression RNAseq (mRNA), and miRNA mature strand expression RNAseq (miRNA). We incorporate DNA methylation measured from the Illumina Infinium HumanMethylation450 BeadChip platform into our analysis, with 90% of the probes from the HumanMethylation27 data present. For gene expression RNAseq, we select the IlluminaHiSeq pancan normalized version. In this modality, values were transformed using $Log2$ mean-normalized per gene across all TCGA patients. Then, only the converted data specific to the cohort of interest was extracted. Moreover, we include miRNA obtained from the IlluminaHiseq system, where values were $Log2$-transformed. More details can be found at the UCSC Xena platform (Goldman et al., 2020).

BLCA is the most common type of bladder cancer, more prevalent in men than women, and can be grouped into low-grade and high-grade cases (The Cancer Genome Atlas Research Network, 2014). LGG is a type of primary brain tumor that includes grades II and III for classification purposes (The Cancer Genome Atlas Research Network, 2015). RCC is the most prevalent type of kidney cancer in adults, which has kidney chromophobe (KICH), kidney clear cell carcinoma (KIRC), and kidney papillary cell carcinoma (KIRP) as the most frequent histological subtypes (Chen et al., 2016).

Table 4.1 Summary of the multiomics data characteristics used in the `HeteroGATomics` experiments.

Dataset	Categories	#Patients	#Original features			#Preprocessed features		
			DNA	mRNA	miRNA	DNA	mRNA	miRNA
BLCA	High-grade: 397, Low-grade: 21	418	485,577	20,530	2,210	7,999	2,373	249
LGG	Grade II: 254, Grade III: 268	522	485,577	20,530	2,157	8,277	1,166	287
RCC	KICH: 65, KIRC: 201, KIRP: 294	560	485,577	20,530	1,847	4,107	2,456	238

4.4.2 Data Preprocessing

After collecting the datasets, we perform four preprocessing steps to clean and prepare the data for machine learning models. First, only patients with available information in all omics modalities are retained. Second, features with missing values in each omics modality are removed to ensure a complete and consistent feature space across all patients (El-Manzalawy et al., 2018). This simplifies the learning process by focusing on observed data and ensures that all features are directly comparable during training. Third, all features in each omics modality are normalized to the range $[0, 1]$ using min-max scaling (You et al., 2020) to ensure comparability across features with different scales (Schulte-Sasse et al., 2021; Wu et al., 2024a), which is required for graph construction in `HeteroGATomics`. This normalization is applied after the *Log*2 normalization performed during data collection to standardize feature ranges while preserving relative differences. Fourth, features with little discriminatory power are eliminated by filtering out those with variance below a specific threshold. The threshold is set at 0.04 for DNA and mRNA, while miRNA is exempt from filtering due to its limited number of features. All datasets in the experiments use this variance threshold. A higher threshold risks removing informative features and reducing diversity, whereas a lower threshold increases the number of features and expands the computational search space, which requires more computational resources. Table 4.1 summarizes the dataset characteristics after completing these phases. To facilitate reproducibility, the processed datasets are available in the GitHub repository.[1]

[1] `https://github.com/SinaTabakhi/HeteroGATomics/tree/main/raw_data`.

4.4.3 Baselines

We compare the performance of the feature selection module of `HeteroGATomics` (that is, $\texttt{HeteroGATomics}_{\text{MAS}}$) with four baseline methods: (1) mutual information (MI) (Theodoridis and Koutroumbas, 2008) quantifies the dependency between two random variables, yielding a non-negative value that is often used to select features with the highest information shared with the target class; (2) recursive feature elimination (RFE) (Guyon et al., 2002) recursively removes the least important features based on an estimator's weights, starting with all features and stopping at the desired number; (3) minimal-redundancy–maximal-relevance (mRMR) (Peng et al., 2005) selects features that have the highest relevance with the target class and are minimally redundant with each other, balancing the trade-off between relevance and redundancy; (4) minimal-redundancy–maximal-relevance multi-view (mRMR-mv) (El-Manzalawy et al., 2018) is an adaption of mRMR to multiomics integration setting. Notably, mRMR-mv is specifically designed for multimodal feature selection within multiomics datasets. To evaluate the performance of MI, RFE, and mRMR, we concatenate the selected features from each modality to serve as the input for a classifier.

Furthermore, we compare the classification performance of `HeteroGATomics` with that of eight multiomics integration methods. We adopt early fusion using five classifiers—*k*-nearest neighbors (KNN) (Theodoridis and Koutroumbas, 2008), multilayer perceptron (MLP) (Theodoridis and Koutroumbas, 2008), random forest (RF) (Ho, 1995), Ridge regression (Ridge) (Hoerl and Kennard, 1970), gradient tree boosting (XGBoost) (Chen and Guestrin, 2016)—concatenating 100 features from each modality selected by MI. We also apply multimodal feature selection using mRMR-mv, a method specifically designed for multiomics data integration, to collectively select 300 features across all modalities based on its selection strategy. Moreover, we use multi-omics graph convolutional networks (MOGONET) (Wang et al., 2021b), a supervised late fusion method for classification tasks, which leverages GCN for omics-specific patient classification and employs VCDN to combine initial predictions from each omics modality into a final label. In addition, we employ the recently developed multiomics method MOSGAT (Wu et al., 2024a) for cancer classification, which leverages specificity-aware GATs for omics-specific learning and a cross-modal attention mechanism

to capture inter-omics associations. For both methods, 100 features from each modality are selected using MI for a fair comparison.

4.4.4 Evaluation Metrics

To evaluate model performance for binary classification tasks, we use six metrics, including area under the receiver operating characteristic curve (AUROC), accuracy, negative predictive values (NPVs), positive predictive values (PPVs), sensitivity, and specificity. For multi-class classification, we use six metrics, including accuracy, macro-averaged F1 score (Macro F1), micro-averaged F1 score (Micro F1), weighted-averaged F1 score (Weighted F1), precision, and recall.

4.4.5 Evaluation Strategies

We evaluate the classification performance on three datasets using stratified 10-fold cross-validation. This approach ensures robust performance evaluation by splitting each dataset into training and test sets while maintaining a balanced class representation. In each fold, nine sets are used for training, further split into training and validation sets at a 9:1 ratio, while the remaining set is used for testing. This process is repeated ten times, ensuring each set is used for testing exactly once. We report the mean and the standard deviation of the evaluation metrics calculated on the test sets across experiments. For the hyperparameter configuration of `HeteroGATomics`, we report the mean of the evaluation metrics calculated on the validation sets. To ensure a fair comparison across different methods, we maintain identical splits for all evaluations.

4.4.6 Implementation Details

`HeteroGATomics` has been developed using Python 3.10 and PyTorch Geometric 2.4.0 (Fey and Lenssen, 2019), incorporating essential functionalities from PyTorch 2.1.0 (Paszke et al., 2019), scikit-learn 1.3.0 (Pedregosa et al., 2011), NumPy 1.26.0 (Harris et al., 2020), pandas 2.1.1 (The pandas development team, 2023), and SciPy 1.11.3 (Virtanen et al., 2020). The

training process for the GAT module and VCDN utilizes the PyTorch Lightning 2.1.3 (Falcon and The PyTorch Lightning team, 2023). For the training of omics-specific encoders using the GAT module and the VCDN, the Adam optimizer (Kingma and Ba, 2015) is employed with the StepLR learning rate scheduler strategy, which reduces the learning rate by a factor of 0.8 every 20 epochs. Each omics-specific encoder comprises a three-layer GAT model with hidden dimensions set to [100, 100, 50], incorporating LeakyReLU nonlinearity with a negative slope of 0.01 after each layer. The decoders for each omics type utilize a single-layer fully connected neural network to map the final 50-neuron hidden layer to the output labels. The omics-specific encoder-decoder pairs is pre-trained for 500 epochs, followed by a full training of the entire architecture for an additional 500 epochs.

For the multimodal feature selection module, $\texttt{HeteroGATomics}_{\text{MAS}}$, the following parameters are used: the maximum number of iterations $T = 50$, the number of agents per omics modality $N_A = 10$, node decay coefficient $\rho_{\mathcal{V}} = 0.1$ in Equation (3.5), edge decay coefficient $\rho_{\mathcal{E}} = 0.1$ in Equation (4.5), omics importance decay coefficient $\gamma = 0.1$ in Equation (3.8), the initial dynamic node score for each node $\tau_u^i(0) = 0.2$, the initial dynamic edge weight for each edge $\tau_{uv}^i(0) = 0.2$, and the state transition rule control parameter $q_0 = 0.8$.

The source code and implementation details of `HeteroGATomics` are available in the GitHub repository.[2]

When performing baseline feature selection methods, we remove five features at each iteration in RFE and employ the recommended hyperparameter values for mRMR-mv as specified in its original paper, ensuring a uniform comparison with $\texttt{HeteroGATomics}_{\text{MAS}}$. We utilize the scikit-feature package (Li et al., 2018) for implementing mRMR, and scikit-learn for implementing MI and RFE.

For classification tasks, we use scikit-learn for KNN, MLP, RF, and Ridge with their default settings, except for the number of iterations in MLP, which is set to 500 epochs. XGBoost is implemented using the XGBoost package (Chen and Guestrin, 2016), with its default configurations. For MOGONET and MOSGAT, we follow the hyperparameter settings recommended in their original publications. To ensure a fair comparison, both

[2] `https://github.com/SinaTabakhi/HeteroGATomics`.

Table 4.2 Classification performance comparison with mean ± standard deviation over 10-fold cross-validation (**best**, second-best).

Dataset	Metric	KNN	MLP	RF	Ridge	XGBoost	mRMR-mv	MOGONET	MOSGAT	HeteroGATomics
BLCA	AUROC	0.779 ± 0.191	0.744 ± 0.194	0.684 ± 0.149	0.720 ± 0.129	0.760 ± 0.175	0.683 ± 0.146	0.884 ± 0.160	**0.961 ± 0.029**	**0.961 ± 0.065**
	Accuracy	0.955 ± 0.027	0.962 ± 0.026	0.955 ± 0.022	0.962 ± 0.016	**0.964 ± 0.016**	0.952 ± 0.018	0.948 ± 0.023	0.955 ± 0.017	**0.964 ± 0.027**
	NPV	0.978 ± 0.020	0.973 ± 0.023	0.968 ± 0.016	0.971 ± 0.018	0.976 ± 0.019	0.968 ± 0.015	0.961 ± 0.019	0.966 ± 0.019	**0.983 ± 0.016**
	PPV	0.517 ± 0.329	0.583 ± 0.417	0.600 ± 0.436	0.650 ± 0.391	0.650 ± 0.369	0.500 ± 0.387	0.292 ± 0.407	0.400 ± 0.382	**0.673 ± 0.360**
	Sensitivity	0.583 ± 0.382	0.500 ± 0.387	0.383 ± 0.299	0.450 ± 0.269	0.533 ± 0.356	0.383 ± 0.299	0.250 ± 0.335	0.350 ± 0.320	**0.667 ± 0.316**
	Specificity	0.975 ± 0.020	0.987 ± 0.017	0.985 ± 0.020	**0.990 ± 0.017**	0.987 ± 0.013	0.982 ± 0.020	0.985 ± 0.023	0.987 ± 0.017	0.980 ± 0.027
LGG	AUROC	0.670 ± 0.052	0.697 ± 0.043	0.704 ± 0.055	0.650 ± 0.049	0.679 ± 0.078	0.687 ± 0.051	0.716 ± 0.050	0.742 ± 0.054	**0.766 ± 0.046**
	Accuracy	0.667 ± 0.053	0.695 ± 0.044	0.703 ± 0.056	0.649 ± 0.049	0.678 ± 0.078	0.686 ± 0.053	0.674 ± 0.060	0.682 ± 0.063	**0.711 ± 0.042**
	NPV	0.630 ± 0.056	0.673 ± 0.050	**0.686 ± 0.071**	0.634 ± 0.048	0.657 ± 0.076	0.670 ± 0.069	0.674 ± 0.073	0.675 ± 0.071	0.675 ± 0.054
	PPV	0.737 ± 0.066	0.726 ± 0.043	0.735 ± 0.058	0.671 ± 0.059	0.715 ± 0.108	0.716 ± 0.047	0.681 ± 0.057	0.692 ± 0.060	**0.783 ± 0.067**
	Sensitivity	0.552 ± 0.116	0.653 ± 0.084	0.664 ± 0.118	0.631 ± 0.073	0.638 ± 0.107	0.645 ± 0.126	**0.689 ± 0.102**	0.683 ± 0.094	0.619 ± 0.120
	Specificity	0.788 ± 0.077	0.740 ± 0.050	0.745 ± 0.083	0.670 ± 0.078	0.721 ± 0.114	0.729 ± 0.072	0.657 ± 0.079	0.682 ± 0.061	**0.808 ± 0.088**
RCC	Accuracy	0.946 ± 0.028	0.954 ± 0.024	0.950 ± 0.026	0.954 ± 0.024	0.955 ± 0.022	0.954 ± 0.021	0.952 ± 0.025	0.950 ± 0.022	**0.961 ± 0.019**
	Macro F1	0.944 ± 0.025	0.953 ± 0.024	0.950 ± 0.020	0.949 ± 0.025	0.955 ± 0.019	0.953 ± 0.018	0.953 ± 0.022	0.950 ± 0.017	**0.957 ± 0.026**
	Micro F1	0.946 ± 0.028	0.954 ± 0.024	0.950 ± 0.026	0.954 ± 0.024	0.955 ± 0.022	0.954 ± 0.021	0.952 ± 0.025	0.950 ± 0.022	**0.961 ± 0.019**
	Weighted F1	0.947 ± 0.028	0.953 ± 0.025	0.950 ± 0.026	0.954 ± 0.024	0.955 ± 0.022	0.954 ± 0.022	0.952 ± 0.026	0.950 ± 0.022	**0.961 ± 0.019**
	Precision	0.949 ± 0.027	0.958 ± 0.021	0.953 ± 0.025	0.956 ± 0.024	0.959 ± 0.020	0.956 ± 0.021	0.956 ± 0.023	0.954 ± 0.021	**0.964 ± 0.019**
	Recall	0.946 ± 0.028	0.954 ± 0.024	0.950 ± 0.026	0.953 ± 0.024	0.955 ± 0.021	0.954 ± 0.021	0.952 ± 0.025	0.950 ± 0.022	**0.961 ± 0.019**

MOGONET and MOSGAT models are run for 500 epochs of pre-training for each omics-specific GCN, and then for an additional 500 epochs for training the entire architecture, similar to the HeteroGATomics setup.

4.4.7 Classification Performance Comparison

We compare the classification performance of HeteroGATomics with multiomics integration methods. Table 4.2 presents the detailed classification comparisons for the BLCA, LGG, and RCC datasets. From Table 4.2, we observe that HeteroGATomics outperforms baseline multiomics integration methods for the binary classification task on BLCA and LGG in terms of AUROC, accuracy, NPV, PPV, sensitivity, and specificity, except for specificity in BLCA, and NPV and sensitivity in LGG. HeteroGATomics still achieves the second-best in NPV for LGG. Furthermore, Table 4.2 demonstrates HeteroGATomics' superiority for the multi-class classification task on RCC, excelling in all the evaluated metrics.

The outstanding performance of HeteroGATomics indicates that the combined power of its $\texttt{HeteroGATomics}_{\text{MAS}}$ module and $\texttt{HeteroGATomics}_{\text{GAT}}$ module trained on generated heterogeneous graphs significantly enhances the capabilities of a deep learning model for multiomics integration.

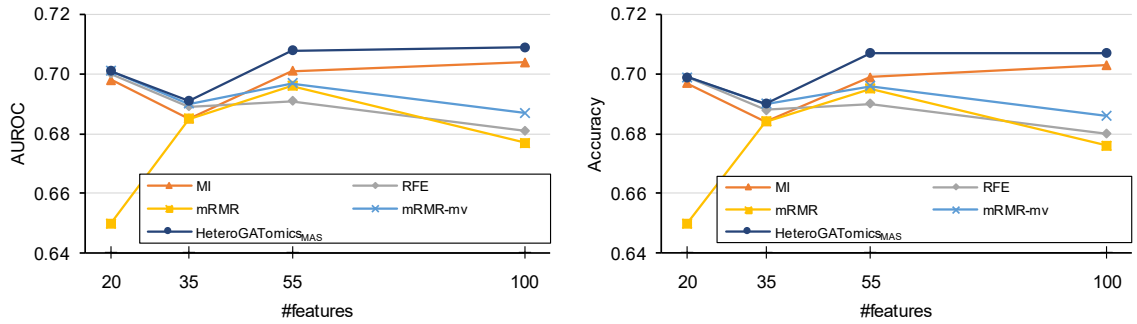

Fig. 4.2 **Performance comparison of feature selection methods for random forest classification on LGG**. The averaged values from 10-fold cross-validation are reported for each metric.

4.4.8 Ablation Studies

We conduct ablation studies to examine the impact of individual modules within `HeteroGATomics` and assess its effectiveness in integrating different modalities. First, we investigate the effectiveness of the standalone `HeteroGATomics`$_{\text{MAS}}$ module. Figure 4.2 presents the classification results of different feature selection methods on LGG using RF, with logarithmically spaced selected feature sizes. `HeteroGATomics`$_{\text{MAS}}$ has consistently outperformed other methods in terms of AUROC and accuracy. MI, a univariate method, achieves the second-highest performance, even against multivariate methods. This suggests that LGG may contain individual features with significant independent predictive power, and the complex interactions identified by multivariate methods may not significantly enhance performance. This finding highlights the effectiveness of `HeteroGATomics`$_{\text{MAS}}$ in identifying discriminative features despite its multivariate nature. Interestingly, both AUROC and accuracy metrics yield remarkably similar trends across all methods, indicating a well-balanced trade-off between true and false positive rates.

Figure 4.3 further validates the generalizability of `HeteroGATomics`$_{\text{MAS}}$ by comparing its performance across three datasets. The figure compares results for 100 selected features from each modality, evaluated with two classifiers, RF and Ridge. On BLCA, `HeteroGATomics`$_{\text{MAS}}$ outperforms all baselines across both evaluation criteria with both classifiers, surpassing the best-reported baseline by 2.8% with RF and 4.2% with Ridge in

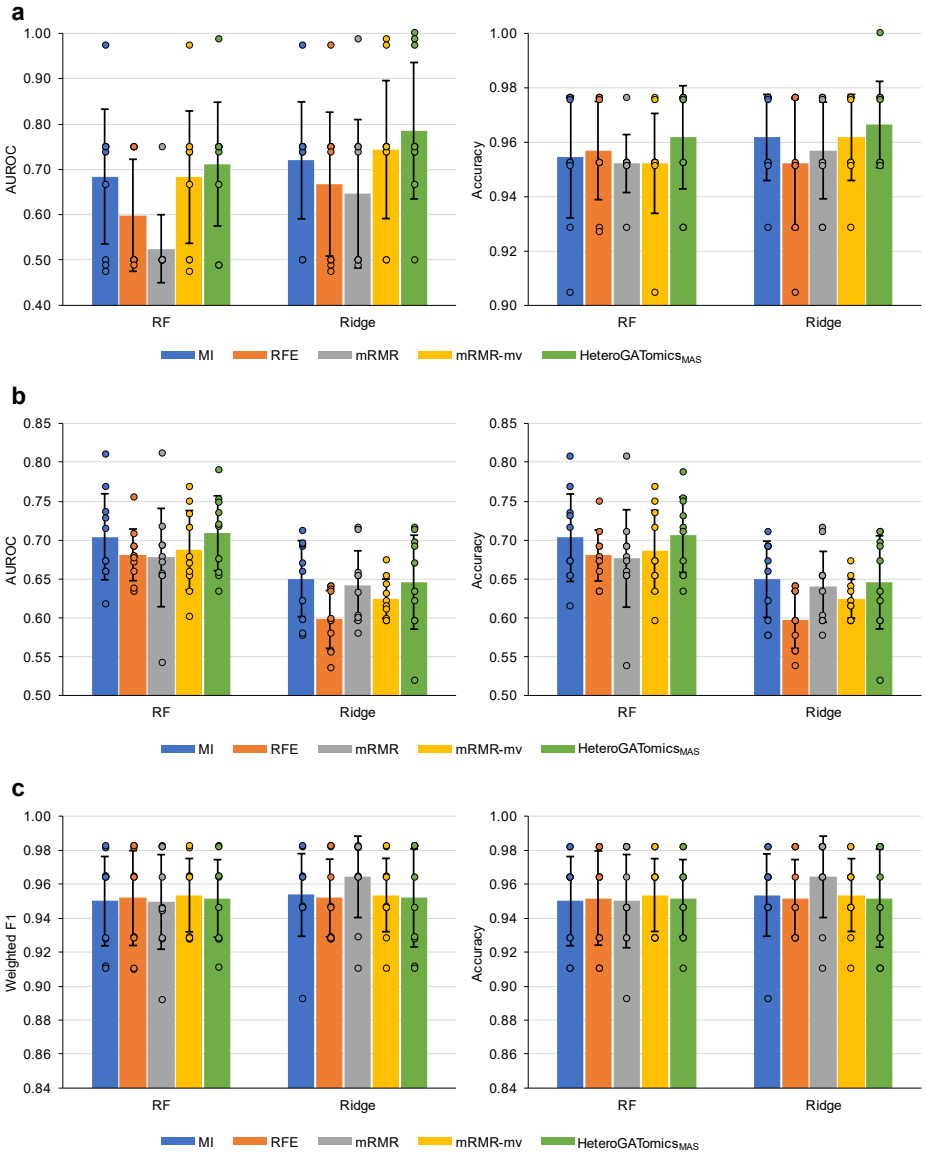

Fig. 4.3 **Performance comparison of feature selection methods for random forest and Ridge classification (mean and standard deviation over 10-fold cross-validation). a**, Results for the BLCA dataset. **b**, Results for the LGG dataset. **c**, Results for the RCC dataset. The results are presented based on 100 selected features for each modality. The vertical bars show the mean, the black lines represent error bars indicating plus/minus one standard deviation, and each dot is a model's performance on each fold. `HeteroGATomics`$_{\text{MAS}}$ denotes the feature selection module within `HeteroGATomics`.

AUROC (Figure 4.3a). For LGG, `HeteroGATomics`$_{\text{MAS}}$ continues to excel in identifying high-quality features, leading to superior evaluation metrics with RF (Figure 4.3b). While

MI slightly outperforms $\texttt{HeteroGATomics}_{\text{MAS}}$ by less than 0.5% in AUROC and accuracy with Ridge, $\texttt{HeteroGATomics}_{\text{MAS}}$ remains the top performer in both metrics when using RF, even surpassing Ridge's results. RCC, known for its relatively straightforward nature in the classification task (Wang et al., 2021b), sees all methods achieving high performance in both metrics with both classifiers (Figure 4.3c). $\texttt{HeteroGATomics}_{\text{MAS}}$ remains competitive on RCC when using RF, with mRMR only outperforming it under Ridge. Additionally, we observe that the efficacy of a given feature selection technique is highly dependent on the dataset. mRMR, for example, is the superior method for RCC, yet it significantly underperforms compared to baselines for BLCA when Ridge is used. These analyses highlight $\texttt{HeteroGATomics}_{\text{MAS}}$'s ability to effectively identify discriminative features across different datasets and classifiers.

Next, we highlight the crucial role of $\texttt{HeteroGATomics}_{\text{GAT}}$ in augmenting the predictive capabilities of HeteroGATomics architecture for classification tasks, beyond what its $\texttt{HeteroGATomics}_{\text{MAS}}$ can achieve. Figure 4.4a compares the whole HeteroGATomics pipeline against $\texttt{HeteroGATomics}_{\text{MAS}}$ on BLCA, LGG, and RCC. We show $\texttt{HeteroGATomics}_{\text{MAS}}$'s performance using KNN, MLP, RF, Ridge, and XGBoost classifiers. Figure 4.4a shows the addition of the heterogeneous GAT module ($\texttt{HeteroGATomics}_{\text{GAT}}$) significantly boosts HeteroGATomics's performance with an improvement of 16.7%, 5.7%, and 0.4% in AUROC on BLCA, LGG, and RCC, respectively. These results highlight the value of the heterogeneous GAT module in HeteroGATomics to enhance its multiomics integration capabilities.

Then, we demonstrate the effectiveness of heterogeneous graphs and their impact on GAT performance. We conduct several modifications: removing the feature similarity network to simulate a homogeneous graph scenario (denoted as Homogeneous), removing static and dynamic edge weights derived from the MAS module as edge attributes (denoted as $\text{Hetero}_{\text{Feature}}$), and excluding static and dynamic node scores derived from the same module as node attributes (denoted as $\text{Hetero}_{\text{Edge}}$). We refer to the full HeteroGATomics architecture, incorporating both node and edge attributes, as $\text{Hetero}_{\text{Feature+Edge}}$. The results of this comparative study on LGG in Figure 4.4b show that HeteroGATomics, when leveraging heterogeneous graphs, outperforms its homogeneous graph counterpart, achieving improve-

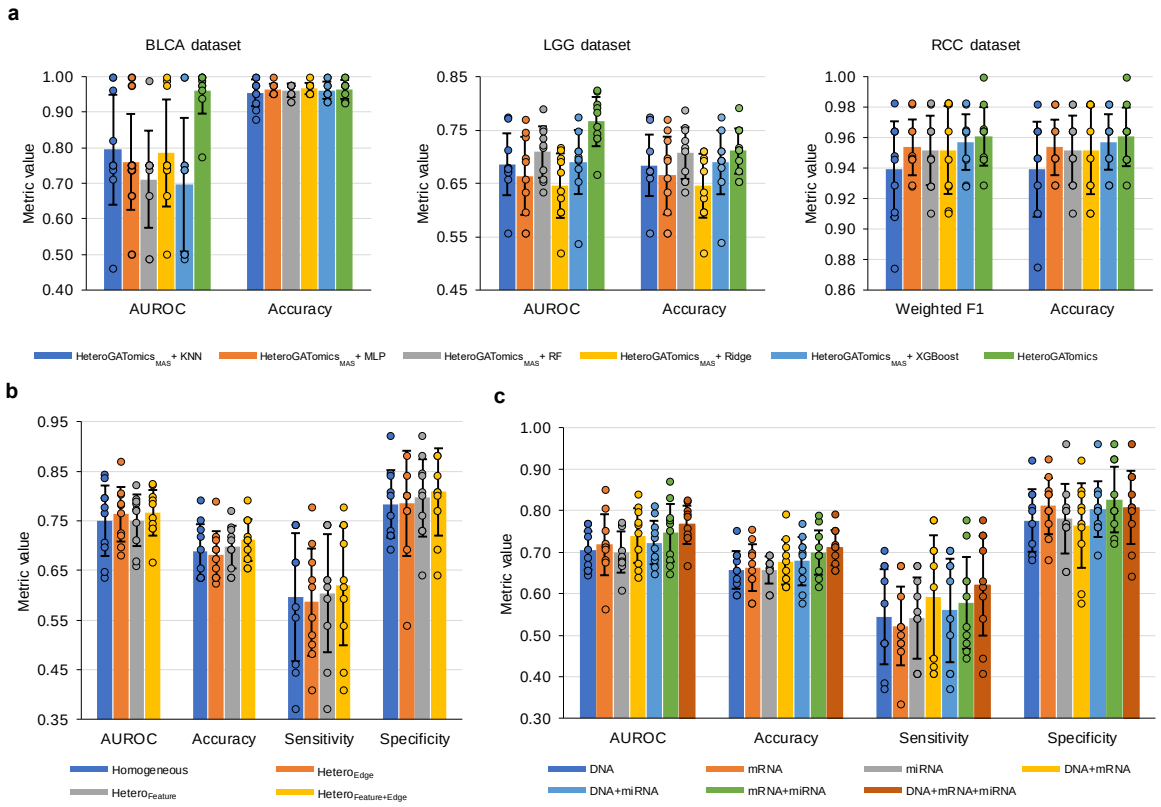

Fig. 4.4 **Ablation studies on** `HeteroGATomics` **performance.** **a**, Evaluation of the feature selection module across five classifiers. $\texttt{HeteroGATomics}_{\text{MAS}}$ + [classifier] denotes the results of the feature selection module within `HeteroGATomics` for a classifier, while `HeteroGATomics` represents the results derived from the entire `HeteroGATomics` architecture. **b**, Evaluation of heterogeneous graph elements using the LGG dataset. Homogeneous refers to `HeteroGATomics` without the feature similarity network, $\text{Hetero}_{\text{Feature}}$ removes edge attributes (static and dynamic edge weights), $\text{Hetero}_{\text{Edge}}$ excludes node attributes (static and dynamic node scores), and $\text{Hetero}_{\text{Feature+Edge}}$ represents the full `HeteroGATomics` setup. **c**, Evaluation of the impact of different omics combinations on `HeteroGATomics` performance using the LGG dataset. DNA, mRNA, and miRNA refer to the single-modality classification performance. Two-modality combinations refer to DNA+mRNA, DNA+miRNA, and mRNA+miRNA, while DNA+mRNA+miRNA refers to the classification performance across three modalities. In each case, 300 features are selected and divided equally among the modalities. The vertical bars show the mean over 10-fold cross-validation, the black lines represent error bars indicating plus/minus one standard deviation, and each dot is a model's performance on each fold.

ments of 2.3%, 1.6%, 2.3%, and 2.4% in accuracy, AUROC, sensitivity, and specificity, respectively. Furthermore, the enhancement in `HeteroGATomics`' performance comes not

only from the feature similarity network but also from the contributions of each individual network element, including feature and edge attributes.

To evaluate the impact of different omics modalities on `HeteroGATomics`' performance and to determine the benefit of integrating multiple omics, we train `HeteroGATomics` using seven modality combinations: individual modalities without the VCDN module (DNA, mRNA, and miRNA), combinations of two modalities (DNA+mRNA, DNA+miRNA, mRNA+miRNA), and all three modalities together (DNA+mRNA+miRNA). The performance of these configurations is assessed on LGG in AUROC, accuracy, sensitivity, and specificity. The results in Figure 4.4c show that integrating additional omics modalities with `HeteroGATomics` enhances its performance across various evaluation metrics, except for specificity. In the context of single-modality learning, mRNA contributes significantly to the performance, particularly in AUROC and accuracy. Notably, integrating all three modalities outperforms single or dual-modality combinations, further highlighting the benefits of multiomics integration via `HeteroGATomics`.

4.4.9 Analysis of Identified Biomarkers for Cancer Diagnosis

To understand the decision-making process within `HeteroGATomics`, we utilize a biomarker importance extraction technique to identify and analyze the key biomarkers that significantly influence the classification results. This process is essential for interpreting the architecture's effectiveness and pinpointing the most informative features serving as cancer biomarkers.

Biomarker identification and interpretation techniques

We leverage an ablation approach commonly used in deep learning-based methods for biomarker selection (Wang et al., 2021b; Setiono and Liu, 1997; Amjad et al., 2022). This approach involves systematically removing each feature to observe its impact on the model's performance. Specifically, for each feature within a given omics, we temporarily remove the feature by setting its value and its attributes (i.e., static and dynamic node scores) to zero in both the feature and patient similarity networks. Then, we evaluate the model's classification performance on the test set without this feature. Features whose removal leads to the

most significant decrease in classification performance are considered top biomarkers. This decrease indicates the feature's substantial impact on the model, highlighting its importance. We apply a 10-fold cross-validation strategy, indicating that the set of input features to the GAT module may vary across different folds. Therefore, we assess the model's classification performance for each feature in every fold. To rank features from each omics modality, we measure the cumulative reduction in classification performance, which is then normalized according to the frequency of their occurrences across all folds. To evaluate the performance of `HeteroGATomics`, we use AUROC for binary classification tasks in BLCA and LGG.

GO enrichment analyses are performed with GOseq in R, using GO biological process (GO BP), molecular function (MF), and cellular component (CC). The top 30 biomarkers for mRNA and DNA omics categories are used as enrichment sets with the remaining 300 biomarkers as the background set. Protein-protein interaction networks are generated in Cytoscape, using string-db (Szklarczyk et al., 2023) and ConsensusPathway (Kamburov and Herwig, 2022) databases for known physical interactions. Only direct two-by-two interactions are used, with an interaction confidence of $>= 7$ for string-db and $>= 9.5$ for ConsensusPathway (reported as high confidence interactions). Known cancer-related genes are fetched from the Cancer Gene Census database (Sondka et al., 2024), OncoKB™ Cancer Gene List (Chakravarty et al., 2017), and the Network Cancer Genome (Repana et al., 2019). Target mRNAs for miRNA biomarkers are inferred from starBase (Li et al., 2014), keeping only targets with at least one CLIP experiment evidence and predicted by at least two miRNA target predictor tools (e.g. miRanda and/or TargetScan).

Analysis of identified biomarkers

Table A.1 in Appendix A presents the 30 most important biomarkers identified by `HeteroGATomics` on BLCA and LGG. RCC is excluded from further biomarker identification, serving only as a proof-of-concept for multi-class classification tasks (Wang et al., 2021b).

Top ranking biomarkers for LGG consist of 16 DNA methylation features, 13 mRNAs, and 1 miRNA. BLCA consists of 14 mRNA features, 11 DNA methylation features, and 5 miRNAs. Gene ontology (GO) enrichment analysis on the top 30 DNA and mRNA biomark-

ers highlights terms related to synaptic and signal transduction for LGG (GO:0097060, p=0.009 and GO:0007165, p=0.012, see Figure A.1a in Appendix A). Belonging to those categories is PTPRA, a protein tyrosine phosphatases with known roles in tumorigenesis (Lv et al., 2023). CHRND and KCNC2, ion channels genes, also belong to both categories. Numerous ion channels are dysregulated in glioma and significantly impact prognosis. Xu et al. (2017) have previously identified KCNC2 as downregulated in malignant gliomas, while CHRND has been identified as a biomarker for treatment and prognosis of head and neck cell carcinomas (Li et al., 2021a).

GO terms related to the top BLCA biomarkers include several developmental processes (e.g. Regionalization, p=0.008, see Figure A.1b in Appendix A). This emphasises top biomarkers that are Transcription Factors (TFs), essential regulators of gene expression: YBX2, HOXB2 and 3, FOXH1, FOXA3, DMRTA2, TEX15 and MAGEA10. These have diverse functions: DNA repair, cell differentiation and migration, cell cycle and organogenesis to cite a few, all relevant in the context of cancer. Among them, TFs of the Fox superfamily FOXH1 and FOXA3 are identified as mRNA biomarkers. FOX family proteins have been shown to be involved in bladder development and cancer progression, for instance FOXA1 expression has been correlated to poor survival (Yamashita et al., 2017). For DNA methylation features, HOXB2 has recently been shown to be upregulated in some subtypes of BLCA and to act as a tumor promoter (Liu et al., 2019). All of the following genes identified by `HeteroGATomics` have been previously flagged as potential biomarkers for BLCA: YBX2 (Yuan et al., 2024), DMRTA2 (Deng et al., 2022), TEX15 (Mantere et al., 2017) and MAGEA10 (Verma et al., 2024).

Notably, the top 5 biomarkers of BLCA are all miRNAs, small RNAs known for regulating gene expression post-transcriptionally by binding target mRNAs and preventing their translation via degradation or translational silencing. As such they play a central role in cell maintenance and in tumorigenesis. Numerous studies have pointed out the importance of miRNAs in bladder cancer and their potential use in cancer diagnosis and prognosis (Das et al., 2023; Sequeira et al., 2023). All 5 miRNAs have been studied in the context of other cancers, for example pancreas and liver (Lee et al., 2023) or breast (Khodadadi-Jamayran

et al., 2018). Three of our five (hsa-mir-1-3p, hsa-mir-708-5p and hsa-mir-16-2-3p), have been previously linked to bladder cancer (Tan et al., 2022; Song et al., 2013; Ware et al., 2022). Zhang et al. (2018a) demonstrated that hsa-mir-1-3p is downregulated in bladder cancer tissues and is able to inhibit cancer proliferation and tumorigenesis in vivo. `HeteroGATomics` is able to not only identify known miRNAs related to bladder cancer but also identify new potential miRNAs, hsa-mir-1976 and hsa-mir-24-3p, involved in oncogenesis.

We further investigate the interaction network of the top biomarkers, including protein-binding partners of protein coding biomarkers and targets of miRNAs biomarkers (Figure 4.5 and Figures A.2 and A.3 in Appendix A). Interestingly, the BLCA miRNA biomarker hsa-mir-708-5p targets four of the protein-coding biomarkers, while hsa-mir-1-3p targets three. This includes partners such as HOXB2 and YBX2, suggesting a possible regulation between them (Figure 4.5a). Examining the protein-protein interactions of the LGG biomarkers shows that RAD9A has numerous protein partners and many are known genes related to cancer (Figure 4.5b and Figure A.3 in Appendix A). RAD9A itself is involved in DNA repair, interacts with several components of the DNA damage response pathways that is targeted for glioblastoma treatment (Ferri et al., 2020). For instance, RAD9A interacts with ATR to mediate DNA repair and several ATR inhibitor drugs have been developed for treating Glioblastoma (Ferri et al., 2020). RAD9A also interacts with HDAC1, a mediator of chromatin compaction, known to be frequently overexpressed in LGG and also targeted for therapy (Cascio et al., 2021). Interestingly, another biomarker, MIDEAS, also interacts with HDAC1, as well as with HDAC2 (Figure 4.5b), forming part of the mitotic deacetylase (MiDAC) complex, which is important for neuronal development (Mondal et al., 2020).

The homeobox TF CUX1 is another important component of LGG. Like MIDEAS, CUX1 is also a target of the miRNA biomarker hsa-mir-363-3p (Figure 4.5b). Moreover, CUX1 has been previously identified as widely expressed in glioma. CUX1 seems to promote tumorigenesis via the Wnt/b-Catenin pathway. Another biomarker, RBMS3, is an RNA binding protein that regulates crucial cellular processes such as transcription, cell apoptosis or cell cycle progression. Its expression correlates with good or poor prognosis depending on cancer type. Ruan et al. (2023) recently discovered that RBMS3 downregulation leads to

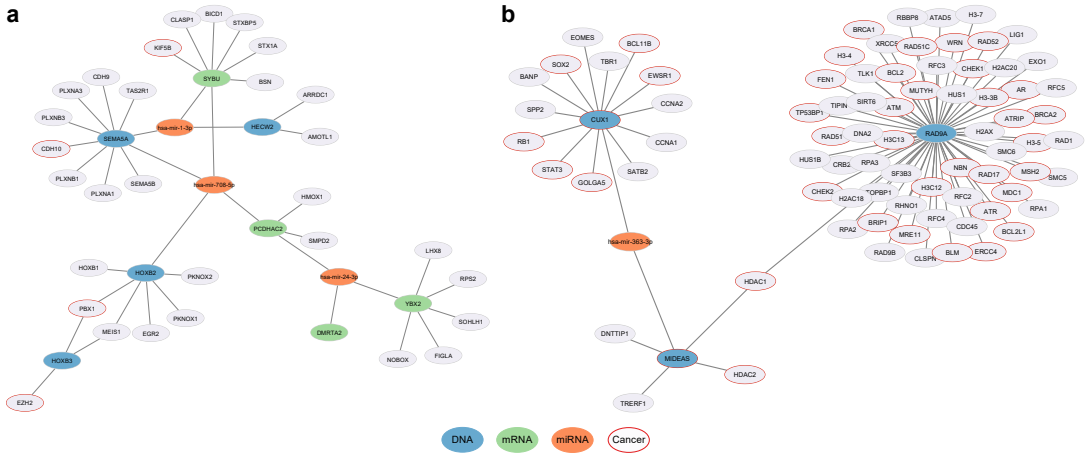

Fig. 4.5 **Known partners of selected top biomarkers. a**, Results for the BLCA dataset. **b**, Results for the LGG dataset. Direct protein-protein interactions are recovered for DNA and mRNA omics. For the miRNA omics, known mRNA targets are recovered from starBase. The different omics categories from which the biomarkers originate are indicated as blue (DNA), green (mRNA) and orange (miRNA). Known cancer-related genes are circled in red.

increased proliferation of glioblastoma cells by interacting with circHECTD1. Subsequently increasing VE-cadherin levels and promoting vasculogenic mimicry, which has a deleterious effect on anti-vascular therapy. `HeteroGATomics` also highlights deletion of BRINP1, also known as DBC1 for deleted in bladder cancer 1. Silencing of, or deletions on, chromosome 9 are known in the context of bladder cancer and include silencing of BRINP1. BRINP1 has been shown to suppress cell cycle progression and to be a candidate tumor supressor (Nishiyama et al., 2001).

4.4.10 Discussion

The development of cancer involves multiple biological layers. The availability of diverse biological data, multiomics, offers detailed insights into various cancer mechanisms. However, integrating multiple heterogeneous omics datasets for cancer diagnosis and biomarker identification remains a significant challenge, especially in small patient cohorts with high-dimensional feature spaces. Our `HeteroGATomics` is a GAT-based method enhanced with heterogeneous graphs to integrate multiomics data for cancer diagnosis. To construct these heterogeneous graphs, we employ a novel dual-view representation leveraging feature and

patient similarity networks. HeteroGATomics utilizes a multimodal feature selection approach using a multi-agent system, effectively addressing the high dimensionality of the feature space.

Experimental results demonstrate that the standalone multimodal feature selection module of HeteroGATomics consistently outperforms baseline feature selection methods and achieves competitive performance in the remaining cases. This suggests that employing an effective multimodal feature selection strategy across multiple modalities can outperform independent feature selection for each modality. Another key innovation of HeteroGATomics is the construction of heterogeneous graphs through our proposed dual-view representation. Leveraging these heterogeneous graphs has significantly enhanced the performance of HeteroGATomics over the feature selection module across all datasets in the experiments, particularly in terms of AUROC. Furthermore, experiments show that, on average, HeteroGATomics outperforms both conventional and state-of-the-art multiomics integration methods in cancer classification tasks across six evaluation metrics.

Ablation studies demonstrate that each component of the heterogeneous graphs positively impacts HeteroGATomics' performance compared to homogeneous graphs. This highlights the importance of leveraging auxiliary information inherent in multiomics datasets for cancer classification and HeteroGATomics' superiority when utilizing this information compared to other methods. Additionally, ablation studies reveal the advantage of training HeteroGATomics with multiple omics over fewer omics, reflecting the complementary information each omics provides to the model. HeteroGATomics is flexible, enabling the integration of additional omics that contribute to understanding cancer progression.

Interpretability is another crucial aspect of this chapter, essential for building trust in HeteroGATomics. We enhance interpretability through a biomarker identification process that explains which genes and features have the most significant impact on the classification of BLCA and LGG cancers. HeteroGATomics successfully predicts known biomarkers in each cohort, including known genes involved in tumorigenesis, diagnosis and/or treatment. We show via interaction networks that several biomarkers are associated and highlight important functions that are or could be targeted for therapy (e.g. DNA damage repair pathwyas in

LGG). HeteroGATomics also identifies genes that were not previously studied in LGG or BLCA, with several known oncogenes in other cancers. Examples of novel therapy target candidates for LGG include ion channels genes CHRND and KCNC2. For BLCA, the top five miRNAs biomarkers can progress miRNA pools used for diagnosis and prognosis. We highlight two network clusters of either interconnected markers, or markers that share common interactors. These may well be highlighting key regulatory processes, where a single process can be disrupted in multiple way to achieve oncogenisity. HeteroGATomics also uncovers non protein coding genes of interest like pseudogenes including long non coding RNAs (lncRNAs, e.g. RNF126P1 and TTTY14 in LGG). Their relevance in cancer has been more recently investigated. For instance dysregulation of lncRNAs has been shown to impact glioma development (Wu et al., 2022).

4.5 Summary

In this chapter, we presented HeteroGATomics, a novel deep learning framework based on GATs for integrating multiomics data in cancer diagnosis and biomarker identification. The model incorporates a multi-agent system for multimodal feature selection, which also provides auxiliary structural information derived from the selected features. Furthermore, we introduced a dual-view representation for constructing heterogeneous graphs that capture both patient-level and feature-level structures, enabling a holistic approach to graph representation learning. Comprehensive experiments on three public cancer multiomics datasets demonstrated the superiority of HeteroGATomics over baseline methods. In addition, ablation studies validated the contribution of each module, and the biomarker identification results highlighted the model's effectiveness in discovering cancer-related biomarkers.

Future work may extend the flexible architecture of HeteroGATomics to incorporate additional modalities beyond biological omics, such as medical imaging and electronic health records. This extension would enable the model to learn more complex relationships across multiple modalities. Moreover, the dual-view representation concept in HeteroGATomics is broadly applicable, extending beyond the specific problems and datasets used in this

study. This approach can be adapted to tabular-structured datasets and applied to other domains beyond cancer genomics, wherever explicit or implicit meaningful relationships among features exist. Finally, `HeteroGATomics`, like many existing methods, currently requires patients to have all omics available and cannot handle missing modalities. Enabling `HeteroGATomics` to learn from partial modalities would further enhance its applicability and performance.

Chapter 5

Multimodal Learning with Missing Modalities

So far in this thesis, we have assumed that all omics modalities are available for each patient. However, a key challenge in learning from multimodal biological data is the presence of missing modalities, where all data from some modalities are missing for some patients. Therefore, in this chapter, we address **Research Question 3**, as outlined in Section 1.2. We propose `MAGNET` (**M**issing-modality-**A**ware **G**raph neural **NET**work) for direct prediction with partial modalities, which introduces a patient–modality multi-head attention mechanism to fuse modality embeddings based on their importance and missingness. `MAGNET`'s complexity increases linearly with the number of modalities while adapting to missing-pattern variability. To generate predictions, `MAGNET` further constructs a patient graph with fused multimodal embeddings as node features and the connectivity determined by the modality missingness, followed by a conventional graph neural network. Experiments on three public multiomics datasets for cancer classification, with real-world instead of artificial missingness, show that `MAGNET` outperforms state-of-the-art fusion methods.

5.1 Introduction

Cancer development is a complex process driven by interactions across multiple molecular layers (Swanton et al., 2024; Yang et al., 2025; Su et al., 2025). To unravel this complexity, cancer research increasingly profiles patients using these molecular modalities, known as multiomics. Each omics modality provides unique value individually while multimodal fusion can offer complementary insights (Karczewski and Snyder, 2018; Acosta et al., 2022). Multimodal machine learning approaches integrate these biological modalities to construct a comprehensive patient profile for improving downstream predictive tasks, such as cancer classification and subtyping (Zitnik et al., 2019; Bi et al., 2025; Cantini et al., 2021).

Despite the effectiveness of multimodal biological data fusion, conventional approaches often assume that all omics modalities are available for each patient (Wang et al., 2014, 2021b). However, missing modalities, characterized by structured missingness where all data from some modalities are missing for some patients, are an unavoidable challenge in biomedical applications (Mitra et al., 2023). For example, some patients may have missing transcriptomic profiles due to sample degradation or insufficient RNA quality, while others may lack proteomic data because of cost constraints or technical limitations (Song et al., 2020; Vitrinel et al., 2019). Therefore, robust and effective multimodal fusion models are important for handling partial modalities.

As discussed in Section 2.4, there are three main approaches to handling missing modalities. We adopt the direct prediction approach, the most recent of the three. The direct prediction approach refers to learning a model that can make predictions directly from the available modalities without explicitly imputing or reconstructing missing ones (Wu et al., 2024b; Yao et al., 2024). Compared to approaches that exclude patients with missing modalities, direct prediction retains all available patients and avoids discarding valuable patient information. Compared to imputation-based approaches, which aim to recover missing modalities before prediction, direct prediction avoids potential error propagation from inaccurate reconstruction and simplifies the overall pipeline. This is particularly important in multiomics settings, where the number of patients is small and modalities are difficult to accurately infer. Therefore, direct prediction provides a more practical and robust solution

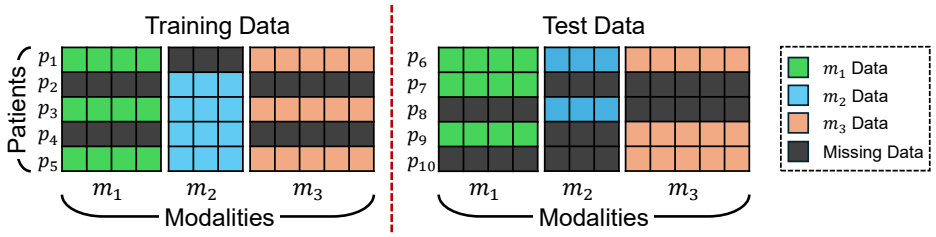

Fig. 5.1 **Different missing-modality patterns across 10 patients with three modalities**. Each colored row within a modality shows a patient's respective data, while a gray row indicates a missing modality for that patient. Moreover, both training and test data can have missing modalities. With three modalities, each patient's data can have $2^3 - 1 = 7$ possible missing patterns.

for real-world clinical scenarios with partial modality availability. However, there are three less-explored challenges for this approach that deserve attention:

- **Challenge 1: scalability with combinatorial missing-modality patterns.** In multi-omics cancer studies, missing modalities arise in various patterns across both training and test data due to data collection constraints (Wu et al., 2024b). For M modalities, there are $2^M - 1$ missing patterns (see Figure 5.1). Direct methods often construct separate sub-models for each pattern (Yun et al., 2024; Yao et al., 2024), or assume missingness occurs in a single modality (Hayat et al., 2022; Yao et al., 2024) or only at test time (Ma et al., 2021; Reza et al., 2024). Such designs do not scale to general missing-modality patterns and therefore present a challenge for cancer profiling.

- **Challenge 2: inflexible modality fusion.** Existing multimodal strategies often assign the same fixed weight to every modality for every patient (Yang et al., 2023), learn a global attention weight for each modality applied to all patients (Ma et al., 2022a; Caruso et al., 2025), or apply attention across patients in a way that treats each patient's fused representation uniformly (Chen et al., 2021b; Keicher et al., 2023). However, in clinical settings, modality importance can vary across patients due to clinical heterogeneity (Yuan et al., 2011) (see Figure 5.2).

- **Challenge 3: artificial missingness evaluation.** A common practice for evaluating missing-modality handling methods involves artificially introducing missingness into

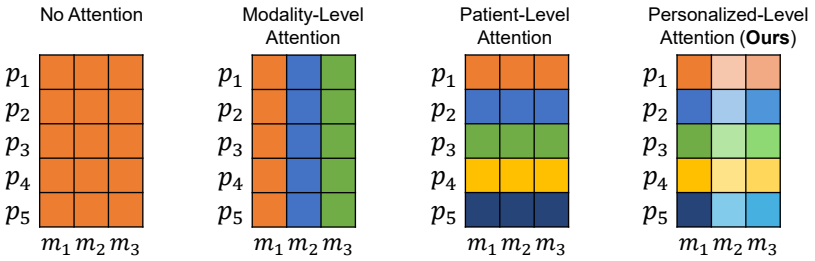

Fig. 5.2 **Comparison of multimodal fusion strategies**. Existing multimodal approaches either use equal weighting across modalities or apply the same attention weights across modalities or patients. However, in real-world clinical settings, modality importance usually differs across patients. For clarity, missing-modality patterns are not shown in these examples.

otherwise complete datasets (Schouten et al., 2018). However, this evaluation approach can oversimplify or misrepresent real-world missingness patterns, particularly when complex, multivariate dependencies are ignored (Mitra et al., 2023). Moreover, it is susceptible to training-time bias, where knowledge about the missing modalities is unintentionally utilized during model training. As a result, this artificial-missingness evaluation approach may fail to reflect real-world missingness challenges.

To address these challenges, in this chapter, we propose `MAGNET` for direct prediction with partial omics modalities. To tackle challenges 1 and 2, `MAGNET` introduces a *patient-modality multi-head attention* (PMMHA) mechanism to fuse modalities into a shared multimodal representation by learning patient-specific modality weights. This design allows a *binary modality mask* to handle missing modalities by zeroing out their contributions. While adding a new modality typically introduces an exponential increase in missing-modality patterns, `MAGNET` requires only one additional learnable attention weight per patient, effectively eliminating the need for separate pattern-specific models. Consequently, `MAGNET` scales linearly with the number of modalities, keeping the model simple and expandable. To preserve inter-patient similarities during fusion, we introduce a *Kullback–Leibler (KL) divergence-based loss* that minimizes the misalignment between patient similarity distributions before and after fusion, even in the presence of missing modalities. After multimodal fusion, `MAGNET` further aligns with real-world clinical practice, where doctors often rely on the insight that patients similar in available modalities are likely to share characteristics in

missing ones (Zhang et al., 2022a). Motivated by this principle, we construct a *patient interaction graph* where nodes represent patients with fused representations as their features, and edges connect patients who share at least one available modality, thereby leveraging missing-modality information. A GNN learns patient representations from this graph to generate final predictions.

To address challenge 3, we evaluate `MAGNET` on multiomics datasets for cancer classification with real-world beyond artificial missingness. The results demonstrate that `MAGNET` consistently outperforms state-of-the-art fusion methods across multiple evaluation metrics.

5.2 Contributions

Our main contributions can be summarized as follows:

- We develop a direct prediction method that operates effectively in the presence of missing modalities and scales linearly with different missing patterns as the number of modalities increases. This is achieved by introducing a patient-modality multi-head attention mechanism that handles diverse missing-modality patterns.
- We propose a new way of constructing the patient interaction graph, which directly incorporates missing-modality patterns into its structure.
- We incorporate KL divergence to preserve relationships among patients, even in the presence of missing modalities.

5.3 Methodology

In this section, we present the architecture of `MAGNET`, which is built upon the initial definitions of supervised multimodal omics learning and multiomics with missing modalities introduced in Section 2.2.5 (Definition 2.2.1 and Definition 2.2.2). Figure 5.3 illustrates the overall architecture of `MAGNET`. Conceptually, `MAGNET` consists of three modules. The

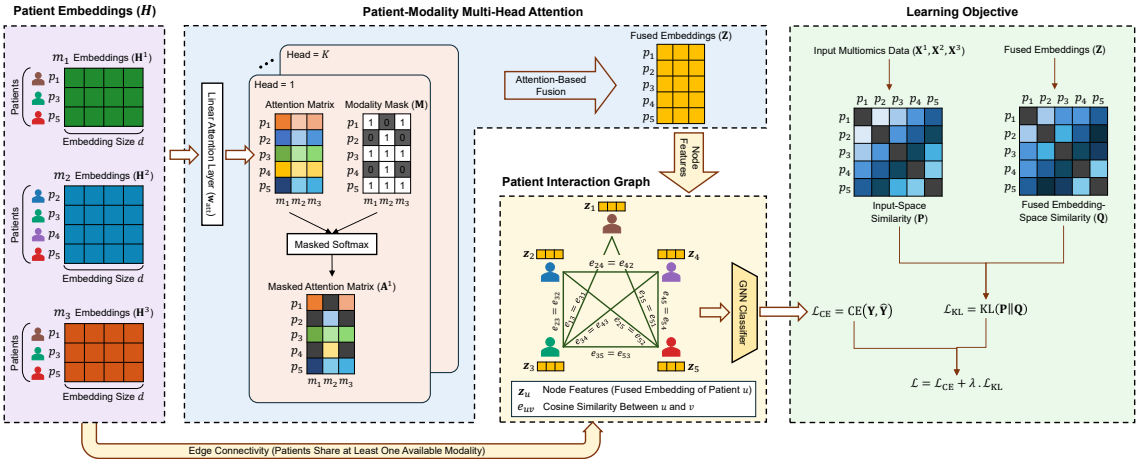

Fig. 5.3 MAGNET **architecture**. MAGNET uses a patient-modality multi-head attention mechanism with learnable parameters ($\mathbf{w}_{\text{att}}$) over patient embeddings (H) and a modality mask ($\mathbf{M}$) to compute patient-specific modality attention weights ($\mathbf{A}^1, \ldots, \mathbf{A}^K$). These weights are used to aggregate patient embeddings into a fused embedding ($\mathbf{Z}$). A patient interaction graph is then constructed, where nodes represent patients with fused embeddings ($\mathbf{z}_u$) as node features, and edges connect patients sharing at least one available modality, with cosine similarity used as the edge feature (e_{uv}). A GNN learns from this graph to perform prediction. The model is optimized using cross-entropy loss ($\mathcal{L}_{\text{CE}}$) for classification and KL-divergence loss ($\mathcal{L}_{\text{KL}}$) to align the similarity distribution of the input space ($\mathbf{P}$) with the fused embedding space ($\mathbf{Q}$).

first module, *Modality-Specific Encoder*, encodes each omics-specific modality into a lower-dimensional representation. The second module, *Patient-Modality Multi-Head Attention Integration*, assigns different weights to each modality for each patient using a binary modality mask and fuses all patient embeddings across modalities into a single unified embedding. The third module, *Patient Graph Construction and GNN Classification*, constructs a graph to represent interactions between patients while accounting for missing modalities. Then, a GNN classifier learns from this constructed graph to make the final prediction. Each module, along with the training strategy, is detailed in the following sections. Furthermore, we discuss the key strengths of MAGNET.

5.3.1 Modality-Specific Encoder

Since each omics modality consists of high-dimensional data with varying dimensionalities, we first transform each modality into a lower-dimensional embedding of the same size

using multilayer perceptron (MLP) encoders. Specifically, for modality i, the corresponding embedding space is constructed as follows:

$$\mathbf{H}^i = \text{MLP}^i(\mathbf{X}^i; \mathbf{W}^i_{\text{MLP}}), \tag{5.1}$$

where $\mathbf{H}^i \in \mathbb{R}^{N \times d}$ is the resulting embedding for N pateints with a fixed dimensionality d, $\text{MLP}^i(\cdot)$ represents the MLP encoder for modality i, and $\mathbf{W}^i_{\text{MLP}}$ represents the learnable parameter of the encoder. This ensures that all modalities are projected into the same size embedding space, facilitating subsequent integration and learning tasks.

5.3.2 Patient-Modality Multi-Head Attention Integration

To address missing modalities and enable the integration of patient embeddings across available modalities, we introduce a PMMHA mechanism. This mechanism calculates attention weights for each modality specific to each patient, enabling selective weighting of modality contributions based on their importance and availability.

In a patient-modality single-head attention (PMSHA) mechanism, the modality-specific embeddings are first stacked to form the tensor $H \in \mathbb{R}^{N \times M \times d}$. A linear transformation is then applied to transform this tensor into higher-level features using a learnable weight matrix $\mathbf{W}_{\text{lin}} \in \mathbb{R}^{d \times d}$.

Next, an attention coefficient matrix is calculated to evaluate the importance of each modality for each patient. To ensure comparability across modalities, the coefficients are normalized using a *masked softmax* operation based on the binary modality mask $\mathbf{M}$. The process is defined as:

$$\mathbf{A} = \text{MaskedSoftmax}(\textsc{Linear}(H; \mathbf{w}_{\text{att}}), \mathbf{M}), \tag{5.2}$$

where $\mathbf{A} \in \mathbb{R}^{N \times M}$ represents the attention coefficient matrix, and $\mathbf{w}_{\text{att}}$ is a learnable parameter. Here, $\text{MaskedSoftmax}(\cdot, \mathbf{M})$ applies a softmax over modalities while assigning $-\infty$ to entries corresponding to missing modalities indicated by the binary mask $\mathbf{M}$, ensuring that missing

modalities receive zero attention weight prior to normalization. Finally, the embeddings for each patient are fused using a weighted sum, where the attention coefficients determine the contribution of each modality. This is performed as:

$$\mathbf{Z} = \sum_{m=1}^{M} \mathbf{a}^m \odot \mathbf{H}^m, \tag{5.3}$$

where $\mathbf{Z} \in \mathbb{R}^{N \times d}$ is the fused embedding for all patients, and $\mathbf{a}^m \in \mathbb{R}^{N \times 1}$ is the attention weights for modality m extracted from $\mathbf{A}$.

To enhance the model's ability to capture diverse patterns in the data, PMMHA allows each head to focus on different aspects of the modality. Therefore, building on PMSHA, the transformed embedding H is divided across K heads, where K is the number of attention heads. The attention process is then applied independently to each head. Specifically, the embedding dimension for each head is $d_h = d/K$, and H is reshaped into $\mathbb{R}^{N \times M \times d_h \times K}$. The normalized attention coefficients for each head k are then calculated as $\mathbf{A}^k$, and the embeddings for each head are fused separately based on these coefficients as $\mathbf{Z}^k$. Finally, the fused embeddings from all heads are concatenated and projected back into the input embedding space as:

$$\mathbf{Z} = \textsc{Concat}(\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^K)\mathbf{W}_{\text{out}}, \tag{5.4}$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{d \times d}$ is a learnable weight matrix, and $\mathbf{Z} \in \mathbb{R}^{N \times d}$ represents the final fused embedding for all patients.

5.3.3 Patient Graph Construction and GNN Classification

With the fused embeddings generated for patients, we aim to make predictions by leveraging both relationships from the input multiomics data and holistic patterns captured in the fused embeddings.

`MAGNET` constructs a patient interaction graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents a set of N nodes corresponding to patients, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents edges connecting patients. Edges are formed using the binary modality mask $\mathbf{M}$, where an edge exists between two patients if

they share at least one available modality. The fused embeddings serve as the initial node features of the graph. Moreover, we assign the cosine similarity between patients as the edge feature, computed from the input multiomics data based only on shared modalities. To ensure the graph focuses on the most relevant patient correlations, we retain only edges whose cosine similarity exceeds a specified threshold β. For nodes without any edge after thresholding, we connect them to their most similar neighbor to preserve graph connectivity.

Given the constructed graph $\mathcal{G}$, `MAGNET` employs a multi-layer GNN, utilizing the graph sample and aggregation (GraphSAGE) operator (Hamilton et al., 2017), to encode the graph. At each layer, the GNN updates a patient's representation by aggregating feature information from its neighbors, allowing the model to incorporate local structural context and capture relationships between patients. This design also enables generating embeddings for unseen patients, as long as their neighborhood information is available (Hamilton et al., 2017; Hamilton, 2020). The GNN stacks multiple GraphSAGE layers, with each layer defined as:

$$\mathbf{m}_{u\leftarrow v}^{(l)} = \mathbf{W}_{\text{msg}}^{(l)} \cdot \textsc{Concat}\left(\mathbf{z}_{v}^{(l)}, e_{uv}\right), \tag{5.5}$$

$$\mathbf{m}_{\mathcal{N}(u)}^{(l)} = \frac{1}{|\mathcal{N}(u)|} \sum_{v\in\mathcal{N}(u)} \mathbf{m}_{u\leftarrow v}^{(l)}, \tag{5.6}$$

$$\mathbf{z}_{u}^{(l+1)} = \textsc{ReLU}\left(\mathbf{W}_{\text{root}}^{(l)}\,\mathbf{z}_{u}^{(l)} \;+\; \mathbf{W}_{\text{agg}}^{(l)}\,\mathbf{m}_{\mathcal{N}(u)}^{(l)}\right), \tag{5.7}$$

$\mathbf{z}_{u}^{(l)}$ is a row of $\mathbf{Z}^{(l)}$, representing the embedding of node u, e_{uv} is the edge feature between nodes u and v, and $\mathbf{W}_{\text{root}}^{(l)}$, $\mathbf{W}_{\text{agg}}^{(l)}$, and $\mathbf{W}_{\text{msg}}^{(l)}$ are learnable weight matrices. In Equation (5.6), $\mathcal{N}(u)$ is the set of directly connected neighbors of node u, and in Equation (5.5), $\textsc{Concat}(\cdot)$ is the concatenation operation. The initial node representations are the fused embeddings obtained from the previous module, given by $\mathbf{Z}^{(0)} = \mathbf{Z}$.

Finally, we use the embedding generated at the last layer L, $\mathbf{Z}^{(L)}$, as input to an MLP decoder to produce the final predictions, defined as $\hat{\mathbf{Y}} = \text{MLP}(\mathbf{Z}^{(L)}; \mathbf{W}_{\text{MLP}})$.

5.3.4 Learning Objective

We optimize the model parameters via backpropagation by minimizing a combined loss that balances prediction accuracy and representation quality. The first component is a *supervised cross-entropy loss*, which encourages the model to correctly classify each patient based on their graph-informed fused embedding. The second component is a *KL divergence-based loss* designed to preserve the structure of patient relationships during fusion. Specifically, it ensures that similarities between patients in the fused embedding space remain consistent with those computed from the input data, despite missing modalities. This encourages the model to maintain meaningful relational information while learning from partially observed data. The supervised cross-entropy loss is defined as:

$$\mathcal{L}_{\text{CE}} = \text{CE}(\mathbf{Y}, \hat{\mathbf{Y}}) = \sum_{(\mathbf{x}_u, \mathbf{y}_u) \in \mathcal{D}_{\text{tr}}} \text{CE}(\mathbf{y}_u, \hat{\mathbf{y}}_u), \tag{5.8}$$

where $\mathcal{D}_{\text{tr}}$ represents all the training patients, and $\text{CE}(\cdot, \cdot)$ is the cross-entropy loss function between true label $\mathbf{y}_u$ and predicted label $\hat{\mathbf{y}}_u$ for pateint u. To preserve patient-level similarity structure in the fused embedding space, we use a KL divergence-based loss:

$$\mathcal{L}_{\text{KL}} = \text{KL}(\mathbf{P} \parallel \mathbf{Q}) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{5.9}$$

where $\mathbf{P}$ is the normalized cosine similarity matrix computed in the input space, and $\mathbf{Q}$ is the normalized similarity matrix in the fused embedding space. Each element q_{ij} is computed using a Student-t kernel (Van der Maaten and Hinton, 2008) as:

$$q_{ij} = \frac{\left(1 + \|\mathbf{z}_i - \mathbf{z}_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|\mathbf{z}_k - \mathbf{z}_l\|^2\right)^{-1}}, \tag{5.10}$$

where $\|\cdot\|^2$ denotes the squared Euclidean distance between the fused embeddings of two patients. The total loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{KL}}, \tag{5.11}$$

Algorithm 5.1 `MAGNET`: Model Training Algorithm

Input

$\mathcal{D} = <(\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^M), \mathbf{Y}>$: multiomics dataset with M omics modality

$\mathbf{M}$: binary modality mask

T_{train}: number of epochs for training

Output

$\hat{\mathbf{Y}} \in \mathbb{R}^{N \times C}$: final predicted labels

Construct patient graph $\mathcal{G}$ with edges formed using $\mathbf{M}$
Set edge feature in $\mathcal{G}$ from patient similarities in $\mathcal{D}$
for $t = 1$ **to** T_{train} **do**
    for $i = 1$ **to** M **do**
        Encode modality i using Equation (5.1)
    end for
    Compute attention $\mathbf{A}$ for each head using Equation (5.2)
    Fuse patient embeddings to obtain $\mathbf{Z}$ using Equation (5.4)
    Initialize graph node features in $\mathcal{G}$ with $\mathbf{Z}$
    Apply GNN to $\mathcal{G}$ to learn embeddings via Equation (5.7)
    Predict labels $\hat{\mathbf{Y}}$ using the MLP decoder
    Compute the cross-entropy loss using Equation (5.8)
    Compute the KL divergence loss using Equation (5.9)
    Update model parameters using the total loss in Equation (5.11)
end for

where λ is a balancing coefficient.

5.3.5 Training Process

We train `MAGNET` using an inductive learning approach, ensuring that patients in the test set are not utilized during the training phase. When constructing the patient graph, all edges connecting patients in the training set to those in the test set are excluded. Moreover, loss computation is performed exclusively on the patients in the training set. A detailed description of `MAGNET` is provided in Algorithm 5.1.

5.3.6 Discussion

Architectural design choices

The design of the MAGNET architecture combines multiple components to enable effective and flexible integration of multiple omics modalities while handling diverse missing-modality patterns in clinical multiomics learning.

MLP encoders are widely used for high-dimensional data as an initial step, as they effectively capture nonlinear relationships between features without requiring additional structural assumptions. This choice is intended to perform independent patient-level feature extraction, projecting raw data into a unified latent space. By avoiding cross-patient interactions at this stage, we ensure that the initial embeddings are purely representative of each individual's biological profile, which simplifies subsequent learning and integration tasks.

PMMHA introduces a patient-specific attention mechanism to assign modality importance for each individual patient through masking, unlike existing methods that use fixed weights across patients or modalities. This is important in clinical settings, where omics modality importance can vary across patients due to clinical heterogeneity. This differs from traditional attention mechanisms that operate between entities of the same type (e.g., patient-to-patient or modality-to-modality) and enables flexible handling of missing-modality patterns in both training and test data.

In clinical multiomics settings, different omics modalities (e.g., DNA vs. mRNA) capture distinct and non-comparable biological aspects, and patients are meaningfully compared only when they share compatible modalities. Accordingly, MAGNET connects patients only when they share at least one modality, ensuring comparisons within a consistent context. This provides a biologically meaningful graph construction strategy that explicitly accounts for missing modalities.

To further ensure the integrity of the fused representations, MAGNET applies a KL divergence loss to patient similarity distributions. This acts as a regularization mechanism, ensuring that even when modalities are missing, the fused latent embeddings remain aligned with the input distributions and preserve the underlying structure of the clinical cohort.

Robustness to missing modalities

Conventional fusion methods often aggregate embeddings across modalities without considering their varying importance. In contrast, MAGNET introduces PMMHA to calculate modality importance for each patient, enabling weighted aggregation of available modalities. To handle diverse missing-modality patterns, we introduce a binary modality mask matrix **M**, which zeros out attention weights for missing modalities. For example, if only one modality is available for a patient, its attention weight becomes one, and its embedding is used. Similarly, if two modalities are available, attention weights are computed for those, and their embeddings are fused accordingly. Even when no modalities are missing, MAGNET performs effectively. This design ensures robust handling of missing-modality patterns in training and testing data.

Linear complexity with modality growth

MAGNET maps each modality to a lower-dimensional space and generates fused multimodal embeddings through the introduced PMMHA mechanism. Adding a new modality requires only an additional encoder for its embeddings and extending **A** with one additional column to represent its importance for each patient. Once fused multimodal embeddings are obtained, GNN operates independently of the number of modalities, relying on the number of patients. Thus, the model achieves linear complexity with respect to M modalities, $\mathcal{O}(M)$, making it efficient for multimodal learning. Details of the experimental execution time analysis are provided in Section 5.4.11.

5.4 Experiments

5.4.1 Data Collection

To evaluate MAGNET, we use three publicly available multiomics datasets from TCGA (Weinstein et al., 2013), obtained via the UCSC Xena tool,[1] which present real-world missingness at varying rates:

- Breast invasive carcinoma (BRCA): This dataset uses the PAM50 classifier, a 50-gene signature, to categorize BRCA into five subtypes based on gene expression: Luminal A, Luminal B, basal-like, HER2-enriched (HER2), and normal-like (Raj-Kumar et al., 2019).
- Bladder urothelial carcinoma (BLCA): This dataset includes bladder cancer cases, classified as low-grade or high-grade for grade classification (The Cancer Genome Atlas Research Network, 2014).
- Ovarian serous cystadenocarcinoma (OV): This dataset contains ovarian cancer samples, divided into long-term (survival time ≥ 3 years) and short-term (survival time < 3 years with 'DECEASED' status) survivors (El-Manzalawy et al., 2018).

We analyze three omics modalities across all datasets: DNA methylation (DNA), gene expression RNAseq (mRNA), and miRNA mature strand expression RNAseq (miRNA). For the DNA modality, we use the Illumina Infinium HumanMethylation27 for BRCA and OV, and the HumanMethylation450 for BLCA. The mRNA modality uses the Illumina HiSeq pancan normalized version, and the miRNA modality includes Illumina HiSeq data. Further details are available on the UCSC Xena platform (Goldman et al., 2020).

5.4.2 Data Preprocessing

To prepare datasets for analysis, we preprocess each omics modality separately. First, we remove features with more than 10% missing values and fill the remaining missing

[1] https://xenabrowser.net/datapages/.

Table 5.1 Summary of the multiomics data characteristics used in the MAGNET experiments.

Dataset	#Patients	Classes	#Patients in Omics (Missing Rate %)			#Selected Features		
			DNA	mRNA	miRNA	DNA	mRNA	miRNA
BRCA	956	Luminal A: 434, Luminal B: 194, Basal-like: 142, Normal-like: 119, HER2-enriched: 67	328 (65.69)	956 (00.00)	584 (38.91)	1,000	1,000	436
BLCA	433	High-grade: 412, Low-grade: 21	431 (00.46)	423 (02.31)	426 (01.62)	1,000	1,000	471
OV	360	Short-term survivors: 184, Long-term survivors: 176	356 (01.11)	184 (48.89)	302 (16.11)	1,000	1,000	448

values using the feature-wise mean. This threshold is chosen to remove features with excessive missingness that may lead to unreliable estimates, while retaining features with limited missing values. While mean imputation may reduce variance and weaken feature relationships, imputing only a small subset of missing values provides a simple strategy that can still maintain the overall structure of the data and has been used in practice (Zhang et al., 2021; Kong et al., 2022). Second, we normalize each feature to the [0, 1] range using min-max scaling (You et al., 2020) to ensure comparability across features with different scales (Schulte-Sasse et al., 2021; Wu et al., 2024a). Third, we exclude low-variance features with limited discriminatory ability, applying modality-specific thresholds: 0.04 for DNA in BRCA and OV, 0.08 for DNA in BLCA, and 0.03 for mRNA across all datasets. These thresholds are chosen to remove near-constant features while retaining sufficient variability, balancing feature reduction and information preservation. No variance filtering is applied to miRNA due to its smaller number of features.

Given the high dimensionality of each omics modality, we further reduce irrelevant and redundant features using ANOVA (Girden, 1992), implemented in scikit-learn. For DNA and mRNA, we retain the top 1,000 selected features to focus on the most discriminative features while keeping the feature space manageable for downstream learning tasks, whereas no feature selection is applied to miRNA due to its limited number of features. This differs from Chapters 3 and 4, where feature selection is part of the proposed methodology. In this chapter, the focus is on the multimodal learning process under missing modalities, and therefore, feature selection is used as a preprocessing step rather than a methodological contribution. Similar feature selection strategies that retain a relatively small subset of

top-ranked features have been adopted in prior multiomics studies (Wang et al., 2021b; Wu et al., 2024a). Table 5.1 presents an overview of the dataset statistics.

5.4.3 Baselines

We compare `MAGNET` with five state-of-the-art multimodal fusion methods. MUSE (Wu et al., 2024b) is a recent method that leverages bipartite graphs for direct prediction. MRGCN (Yang et al., 2023) uses an encoder-decoder framework based on GCNs for direct predictions with missing modalities. M3Care (Zhang et al., 2022a) is an imputation-based method that reconstructs missing modalities in the latent space by leveraging similarity between patients. MOGONET (Wang et al., 2021b) is a supervised multiomics integration method based on GCN that requires complete modalities. To address missing modalities for MOGONET, we apply two imputation strategies, zero imputation and k-nearest neighbor (kNN), and refer to them as MOGONET-Zero and MOGONET-kNN. We set $k = 10$ due to the limited number of patients in the multiomics datasets.

5.4.4 Evaluation Metrics

We evaluate the performance of fusion methods on the BLCA and OV datasets for binary classification using accuracy, area under the precision-recall curve (AUPRC), area under the receiver operating characteristic curve (AUROC), and Matthews correlation coefficient (MCC). For multi-class classification on the BRCA dataset, we assess performance using accuracy, macro-averaged F1 score (Macro F1), weighted-averaged F1 score (Weighted F1), and MCC.

5.4.5 Evaluation Strategies

To evaluate classification performance, we divide each dataset into matched data (patients with all modalities available) and unmatched data (patients with missing modalities). Each subset is further split into training, validation, and test sets with a ratio of 7:1:2, ensuring diverse missing patterns in training and test data. Matched and unmatched subsets are then

combined to form the final training, validation, and test sets. For hyperparameter tuning, models are trained on the training set, and the best parameters are selected by monitoring performance on the validation set. Due to the limited sample size, the training and validation sets are combined after hyperparameter tuning to form the final training set. All methods are trained on this combined training set, and performance metrics are evaluated on the test set. To ensure robustness, all evaluations are performed five times, with results reported as the mean and standard deviation of the metrics. For consistency, the same data splits are used across all methods.

5.4.6 Implementation Details

We implement `MAGNET` in Python 3.10 using PyTorch 2.1.0 (Paszke et al., 2019) and PyTorch Geometric 2.4.0 (Fey and Lenssen, 2019). The Adam optimizer (Kingma and Ba, 2015) is used for training, with a step decay learning rate scheduler that reduces the learning rate by a factor of 0.8 every 20 epochs. We fix the number of layers in the MLP encoder, GNN, and MLP decoder to two across all datasets. The λ parameter in Equation (5.11) is set to 0.1. All models are trained for 200 epochs, with the batch size set to the total number of patients in each dataset due to the limited number of patients. For MOGONET, which includes both a pretraining and training phase, we allocate 100 epochs for pretraining and 100 epochs for training. All baselines are implemented using an inductive learning approach, where test sets are excluded during the training phase. All experiments are run on an Ubuntu 24.04 machine with an NVIDIA GeForce RTX 4090 GPU. The source code of `MAGNET` is publicly available.[2]

To ensure a fair comparison, we perform hyperparameter tuning for `MAGNET` and all baseline methods using the Ray Tune library (Liaw et al., 2018). We tune key hyperparameters that significantly influence model performance. For all methods, we run 100 trials with the Asynchronous Successive Halving Algorithm scheduler (Li et al., 2020) to prioritize promising configurations. Each trial runs for up to 100 epochs, with training on the training set and evaluation on the validation set. To ensure robustness across data splits, we repeat

[2] `https://github.com/SinaTabakhi/MAGNET`.

Table 5.2 Classification performance comparison on the BRCA, BLCA, and OV datasets, reported as the mean ± standard deviation over five independent runs (**best**, second-best).

Dataset	Metric	MOGONET-Zero	MOGONET-*k*NN	MRGCN	M3Care	MUSE	MAGNET
BRCA	Accuracy (↑)	0.795±0.025	0.847±0.015	0.844±0.009	0.899±0.016	0.895±0.021	**0.918±0.012**
	Macro F1 (↑)	0.638±0.062	0.794±0.035	0.826±0.014	0.872±0.018	0.880±0.014	**0.902±0.019**
	Weighted F1 (↑)	0.758±0.036	0.838±0.018	0.842±0.008	0.898±0.015	0.894±0.022	**0.917±0.011**
	MCC (↑)	0.706±0.035	0.781±0.022	0.778±0.012	0.858±0.022	0.851±0.031	**0.884±0.016**
BLCA	Accuracy (↑)	0.952±0.005	0.952±0.005	0.955±0.000	0.968±0.011	0.966±0.014	**0.970±0.006**
	AUPRC (↑)	0.652±0.106	0.688±0.098	0.617±0.086	0.686±0.114	0.653±0.085	**0.724±0.101**
	AUROC (↑)	0.898±0.142	0.902±0.162	0.944±0.033	0.949±0.051	0.939±0.046	**0.956±0.025**
	MCC (↑)	-0.005±0.009	-0.005±0.009	0.000±0.000	0.597±0.164	0.509±0.294	**0.642±0.072**
OV	Accuracy (↑)	0.581±0.027	0.573±0.051	0.597±0.014	0.597±0.031	0.608±0.036	**0.614±0.052**
	AUPRC (↑)	0.630±0.030	0.594±0.044	**0.646±0.022**	0.607±0.051	0.628±0.078	**0.646±0.046**
	AUROC (↑)	0.621±0.037	0.603±0.062	**0.655±0.025**	0.630±0.040	0.652±0.067	0.652±0.056
	MCC (↑)	0.199±0.071	0.126±0.138	0.201±0.029	0.199±0.060	0.225±0.073	**0.228±0.104**

each configuration five times and use the average performance for selection. Hyperparameter ranges are similar across methods unless otherwise specified in their original papers, in which case we adopt the recommended ranges. Table B.1 in Appendix B lists the search spaces and the best-performing values for each dataset and method.

5.4.7 Classification Performance Comparison

Table 5.2 presents the classification performance of fusion methods on the BRCA, BLCA, and OV datasets. MAGNET consistently outperforms baseline multimodal fusion methods across all datasets and evaluation metrics, except for AUROC on OV, where it achieves the second-best result. Below, we provide detailed observations for each dataset.

Results on BRCA

MAGNET performs better than the best-performing direct prediction counterparts, MRGCN and MUSE, with improvements of 2.3% in accuracy, 2.2% in Macro F1, 2.3% in Weighted F1, and 3.88% in MCC. The weaker performance of MUSE may be due to overlapping feature distributions among certain BRCA classes, as it does not explicitly model direct patient-to-patient connections. Similarly, MRGCN fuses patient embeddings across modalities by assigning equal contributions, which may overlook the varying importance of different modalities. Moreover, both methods may struggle to handle severe modality missingness,

as their design assumptions might not effectively capture complex missing patterns. This limitation is further reflected in the performance of MOGONET-Zero, which shows the worst results across all metrics, suggesting that simple zero imputation is ineffective under such missingness.

Results on BLCA

MAGNET maintains its superior performance, achieving a 3.6% improvement in AUPRC and a 7.54% relative improvement in MCC over the best baseline, a particularly notable gain given the dataset's class imbalance. In contrast, MOGONET with imputation yields the worst results across nearly all metrics for BLCA. This may be due to BLCA's relatively low rate of missing modalities combined with a highly imbalanced class distribution, where even minor imputations can introduce errors that especially affect the already limited minority class.

Results on OV

On this well-balanced dataset, MAGNET demonstrates strong overall classification performance, achieving the highest accuracy, AUPRC, and MCC. However, its focus on correctly identifying the positive class may slightly affect ranking performance across all thresholds, leading to a marginally lower AUROC. Despite this, the overall results confirm the robustness of the method. Notably, while M3Care performs well on BRCA and BLCA, it ranks among the worst on OV, highlighting the difficulty for baselines to generalize across datasets with varying characteristics.

MAGNET's superior results highlight the effectiveness of its PMMHA mechanism and graph-based modeling, enhancing multimodal fusion in the presence of missing modalities.

5.4.8 Analysis of Learned Representations

To further evaluate the effectiveness of MAGNET, we measure the separability of patient representations using two clustering metrics: Silhouette Score (SS) (Rousseeuw, 1987), where higher values indicate better-defined clusters, and Davies-Bouldin (DB) index (Davies

Table 5.3 Separability evaluation of patient representations on test data using Silhouette Score (SS) and Davies-Bouldin (DB) index (**best**, second-best).

Method	BRCA		BLCA		OV	
	SS (↑)	DB (↓)	SS (↑)	DB (↓)	SS (↑)	DB (↓)
MOGONET-Zero	0.214	1.447	0.836	0.741	0.024	4.491
MOGONET-*k*NN	0.354	1.078	**0.870**	0.757	0.034	4.395
MRGCN	0.145	1.658	0.261	0.700	0.006	8.808
M3Care	0.437	0.826	0.330	0.670	0.040	4.551
MUSE	0.408	0.879	0.295	0.793	0.047	**3.900**
MAGNET	**0.440**	**0.789**	0.578	**0.642**	**0.048**	4.433

and Bouldin, 1979), where lower values reflect better class separation. We extract the representations from the last layer of each trained model before generating predictions. Table 5.3 presents the results on the test data for three datasets.

On BRCA, MAGNET improves SS and reduces DB, demonstrating its ability to separate classes effectively in a multiclass classification task. On BLCA, an imbalanced dataset, MAGNET achieves the lowest DB index, indicating strong global class separation. However, its SS is moderate, likely due to minority class patients being closer to decision boundaries. In contrast, imputation-based methods yield a higher SS by estimating missing values based on dominant patterns, which results in more compact clusters and improved local cohesion. On OV, a well-balanced dataset, MAGNET achieves the highest SS, highlighting well-clustered patient representations. However, its DB index remains average, likely because balanced class distributions result in uniform cluster compactness, reducing the inter-cluster contrast that DB emphasizes.

These results show MAGNET's ability to enhance class separability across datasets while balancing local cohesion and global structure in learned representations.

5.4.9 Studies on Simulated Missing-Modality Scenarios

To further evaluate the impact of missing modalities, we use a publicly available simulated multiomics cancer dataset generated by the InterSim CRAN package, consisting of 500 samples across 15 clusters of varying sizes, reflecting realistic clinical scenarios. The

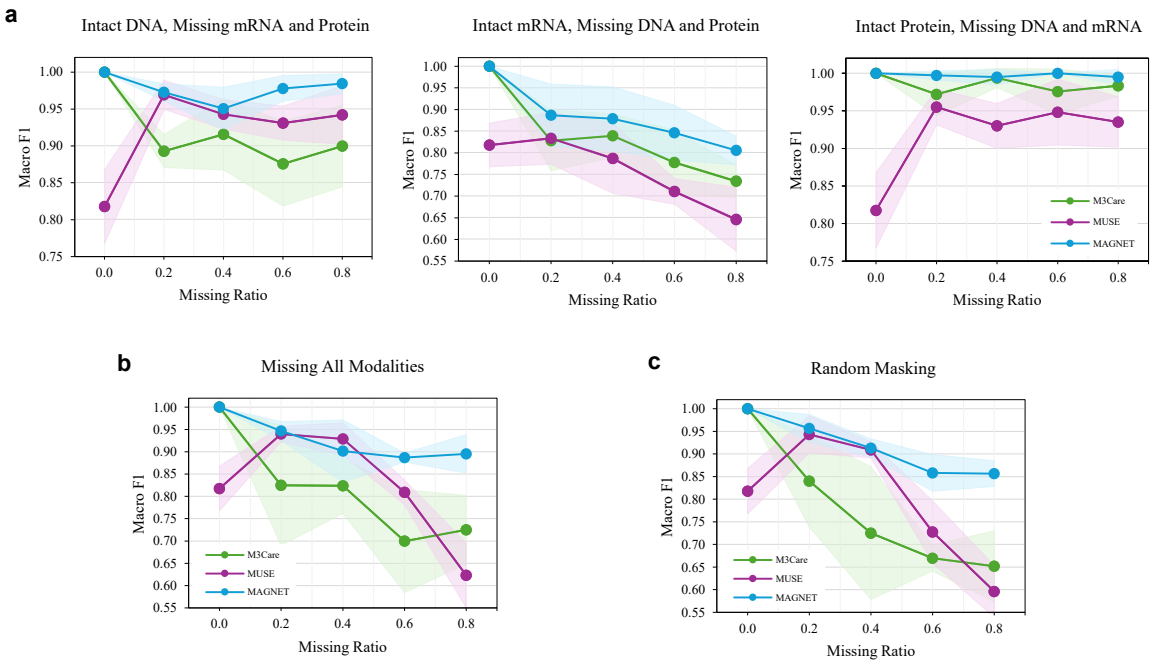

Fig. 5.4 **Effect of varying missingness ratios on simulated cancer data using Macro F1, averaged over five independent runs per ratio. a**, One modality remains intact while the other two are uniformly subsampled. **b**, A subset of patients is shared across all modalities, with the rest uniquely assigned to individual modalities. **c**, Modalities are randomly masked with different probabilities.

dataset includes three omics modalities: DNA methylation, mRNA expression, and protein expression. The data is available on Zenodo (Ma, 2024), and further details about the data generation process can be found in the original publication (Ma et al., 2025).

To prepare the dataset for analysis, we normalize each feature to the [0, 1] range using min-max scaling, applied separately to each omics modality. The data is then split into training and test sets with a ratio of 8:2. We compare `MAGNET` with two top-performing baselines: M3Care and MUSE. We evaluate performance under three missing modality scenarios, and the results are presented in Figure 5.4.

In the first scenario, we keep one omics modality intact while uniformly subsampling the other two at missingness ratios ranging from 0.2 to 0.8, with the zero-missingness case representing the complete dataset (Figure 5.4a). The results show that `MAGNET` consistently outperforms the baselines across all missingness levels. Moreover, we observe that when DNA or protein modalities are kept intact, performance remains relatively stable, indicat-

ing that they are more informative. In contrast, when mRNA is the only intact modality, performance drops significantly, suggesting it is less predictive on its own.

In the second scenario, we retain a subset of patients that are shared across all three modalities and evenly assign the remaining patients to individual modalities. The missingness ratio varies from 0.2 to 0.8 (Figure 5.4b). We observe that MAGNET consistently achieves the highest performance across most settings. Notably, as the missingness ratio increases, the performance gap between MAGNET and the baseline methods also grows, highlighting the model's robustness to missing modality information.

In the third scenario, we adopt a more complex setup where no omics modality is kept fully intact. Modalities are randomly masked with probabilities ranging from 0.2 to 0.8 (Figure 5.4c). Performance results show that MAGNET consistently outperforms the baselines across all missingness levels. We also observe that as the masking probability increases, the performance gap between MAGNET and the baselines gets larger. For example, under severe missingness (80%), MAGNET improves performance by around 20% and 26% compared to M3Care and MUSE, respectively, demonstrating strong resilience to high levels of missing modalities. We also observe that M3Care, as an imputation-based method, experiences a much larger performance drop than the other methods as the missing probability increases. This may be because imputation-based methods rely on reconstructing missing data, which becomes increasingly difficult and error-prone as more information is missing.

5.4.10 Ablation Studies

We conduct ablation studies to evaluate the impact of individual components in MAGNET and assess its effectiveness.

Classification performance analysis

We evaluate the contribution of each component in MAGNET through seven modifications: (A1) removing the PMMHA mechanism and assigning equal contributions to modalities, (A2) removing the GNN architecture and applying the MLP decoder on the fused embedding, (A3) removing the edge feature from the patient interaction graph, (A4) removing the

Table 5.4 Ablation study on the impact of each individual component of MAGNET (**best**, second-best).

ID	Method	BRCA	BLCA	OV
A1	MAGNET w/o PMMHA	$0.905_{\pm 0.016}$	$0.681_{\pm 0.064}$	$0.600_{\pm 0.049}$
A2	MAGNET w/o GNN	$0.907_{\pm 0.019}$	$0.675_{\pm 0.068}$	$0.600_{\pm 0.041}$
A3	MAGNET w/o Edge Feature	$0.910_{\pm 0.015}$	$0.719_{\pm 0.135}$	$0.600_{\pm 0.064}$
A4	MAGNET w/o KL Loss	$0.911_{\pm 0.013}$	$0.686_{\pm 0.104}$	$0.592_{\pm 0.071}$
A5	MAGNET w/ GAT	$0.711_{\pm 0.083}$	$0.606_{\pm 0.127}$	$0.551_{\pm 0.029}$
A6	MAGNET w/ GCN	$0.823_{\pm 0.002}$	$0.609_{\pm 0.110}$	$0.567_{\pm 0.041}$
A7	MAGNET w/ GIN	$0.560_{\pm 0.173}$	$0.490_{\pm 0.157}$	$0.529_{\pm 0.076}$
-	MAGNET	$\mathbf{0.918_{\pm 0.012}}$	$\mathbf{0.724_{\pm 0.101}}$	$\mathbf{0.614_{\pm 0.052}}$

KL loss and using only the cross-entropy loss, and (A5)-(A7) replacing the GNN with a GAT (Veličković et al., 2018), GCN (Kipf and Welling, 2017), and GIN (Xu et al., 2019), respectively. Table 5.4 presents the results. We observe a significant performance drop when the PMMHA mechanism and GNN are removed, indicating that simple aggregation is insufficient for effective fusion in the presence of missing modalities, and that graph-based interactions are essential for capturing relational structure among patients. This is further supported by (A5)-(A7), where replacing the original GNN leads to a consistent drop, highlighting the suitability of GraphSAGE with the edge feature. Overall, these results validate the effectiveness of MAGNET.

Learned representation analysis

We evaluate eight types of representations used in the MAGNET architecture by measuring the separability of patient representations in the test data using SS and DB index: three input modalities, their learned representations from omics-specific encoders, the fused representation from PMMHA, and the final-layer representation produced by the GNN. Table 5.5 presents the results, showing that in all cases across the three datasets, the final-layer representation learned by MAGNET achieves the highest separability. Moreover, the representation learned from PMMHA alone achieves nearly the second-best result across all

Table 5.5 Separability evaluation of patient representations learned from MAGNET modules on test data using Silhouette Score (SS) and Davies-Bouldin (DB) index (**best**, second-best).

Module	BRCA		BLCA		OV	
	SS (↑)	DB (↓)	SS (↑)	DB (↓)	SS (↑)	DB (↓)
Input DNA	0.035	2.635	0.231	0.987	0.002	7.833
Input mRNA	0.105	2.276	0.120	1.531	0.006	5.264
Input miRNA	0.046	3.307	0.024	2.331	0.005	6.961
DNA Embedding	0.017	2.522	0.264	0.977	0.002	7.807
mRNA Embedding	0.246	1.170	0.188	1.284	0.018	4.618
miRNA Embedding	0.018	3.571	0.004	2.421	0.006	6.897
PMMHA Embedding	0.295	0.983	0.345	0.865	0.009	6.585
GNN Embedding	**0.440**	**0.789**	**0.578**	**0.642**	**0.048**	**4.433**

datasets compared to individual omics representations, suggesting that each added component in MAGNET contributes to producing more informative representations.

To further validate the effectiveness of fusing input omics modalities using MAGNET, we visualize the input omics modalities and the representations learned by the last layer of the GNN module in MAGNET using uniform manifold approximation and projection (UMAP). Figure 5.5 shows the UMAP visualizations for the BRCA dataset, which contains a larger number of patients, providing better clarity compared to the other two datasets. We observe that among the individual omics modalities, mRNA demonstrates greater class separability on both the training and test data, whereas DNA and miRNA provide limited separability, with patients within the same classes overlapping significantly. Specifically, for the mRNA modality, while patients within the Normal-like and Basal-like classes are relatively well separated, there is some overlap between patients in the Luminal A and Luminal B classes. In contrast, MAGNET effectively integrates these modalities, enhancing class separability. The fused representation not only distinguishes patients with similar classes more clearly but also achieves better separation between the Luminal A and Luminal B classes compared to the mRNA modality alone.

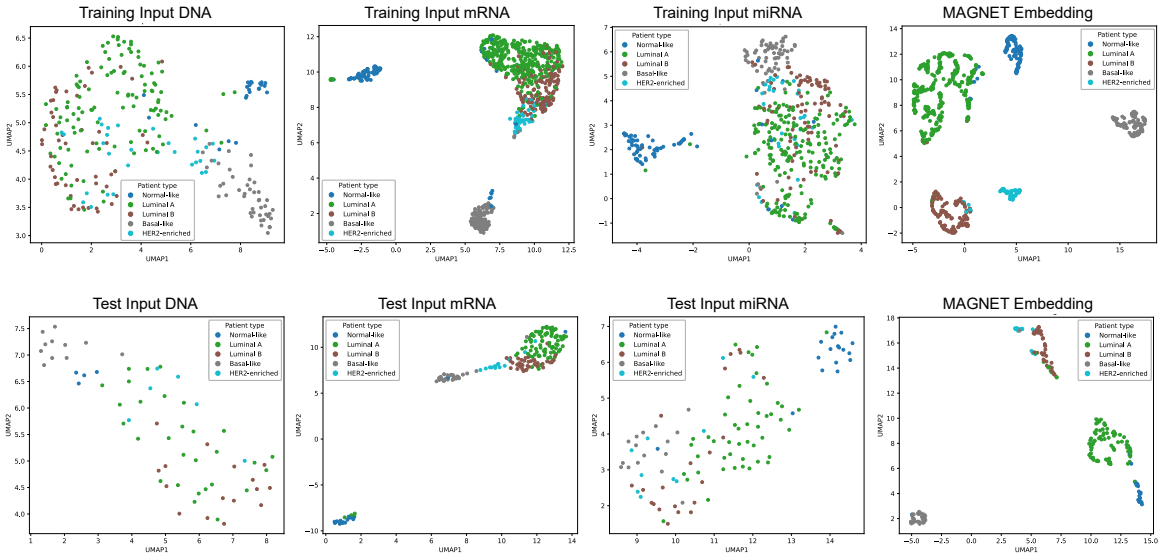

Fig. 5.5 **UMAP visualization of the training and test data from the BRCA dataset.** For each omics modality, UMAP is generated from the input data, while for `MAGNET`, it shows the patient representations learned by the GNN module. **Top:** Training data visualization. **Bottom:** Test data visualization.

Impact of multiomics data integration

There are two key reasons for integrating multiple omics modalities. First, relying on a single modality limits prediction for patients missing that modality. For example, as shown in Table 5.1, the BRCA dataset has approximately 66% missing DNA data. Using only DNA would exclude over half of the patients, highlighting the need to leverage other available modalities. Second, additional modalities can provide complementary information that enhances model performance beyond what a single modality can achieve. Although whether multiomics integration consistently improves performance remains an open question, our focus is on scenarios where integration offers improvements over single-modality models.

To explore this, we conduct an additional ablation study on the BRCA dataset. Specifically, we select patients who have data available for all three modalities, ensuring that when one or two modalities are removed, the remaining modalities are still present for every patient. We train `MAGNET` using seven different modality combinations: individual modalities, pairwise combinations of two modalities, and the full set of three modalities. When removing a modality, we mask it for the corresponding patients, treating it as a missing modality. The

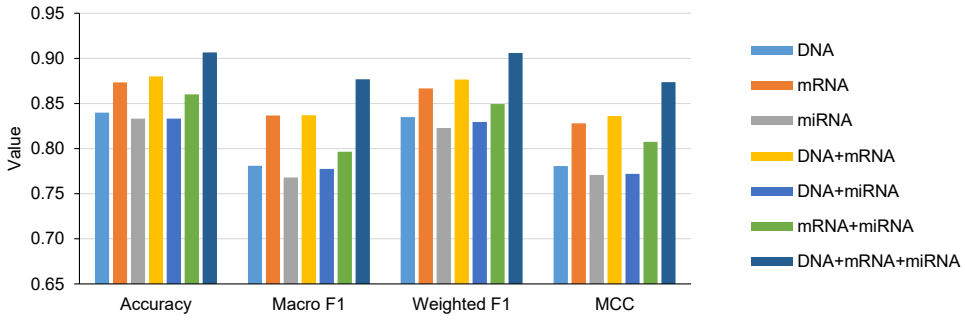

Fig. 5.6 **Performance of MAGNET, averaged over five runs, across different combinations of omics modalities on the BRCA dataset.**

results of this experiment are shown in Figure 5.6. The results show that although mRNA alone has the strongest predictive power among the three modalities, combining it with DNA further improves performance. Moreover, using all three modalities together yields the best overall results across all evaluation metrics.

5.4.11 Execution Time Analysis

We first evaluate the execution time versus classification performance of all methods during training on the BRCA, BLCA, and OV datasets. The results, averaged over five runs, are presented in Figure 5.7. MAGNET achieves the best prediction performance among all methods, although its execution time is relatively higher than some baselines. This demonstrates that MAGNET offers a favorable trade-off between predictive performance and computational efficiency. Among the baselines, MAGNET execution time is slightly higher than that of the recent method MUSE. In contrast, M3Care as an imputation-based method shows the worst computational time which is almost double that of MAGNET. This highlights the high computational cost of imputation-based approaches. The lightweight computational time of MOGONET-Zero and MOGONET-kNN is attributed to the fact that their imputation is a simple procedure performed before training. However, this simplicity comes at the cost of significantly worse classification performance in almost all cases.

We next present the execution time of MAGNET with an increasing number of input modalities. To conduct this experiment, we retain only patients with all modalities available,

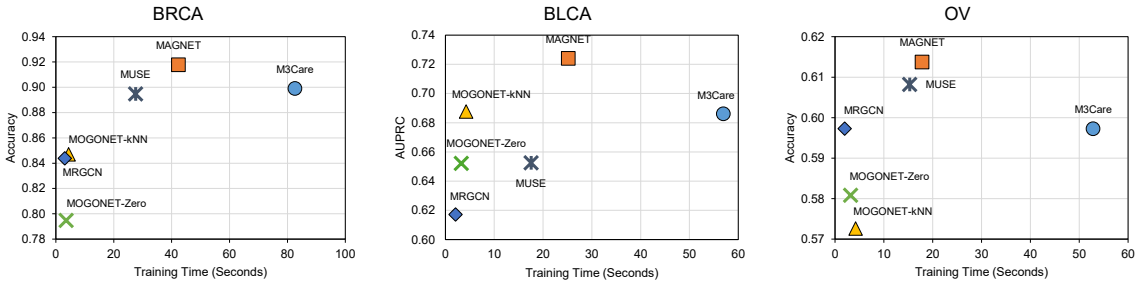

Fig. 5.7 **Comparison of classification performance and execution time, averaged over five runs.**

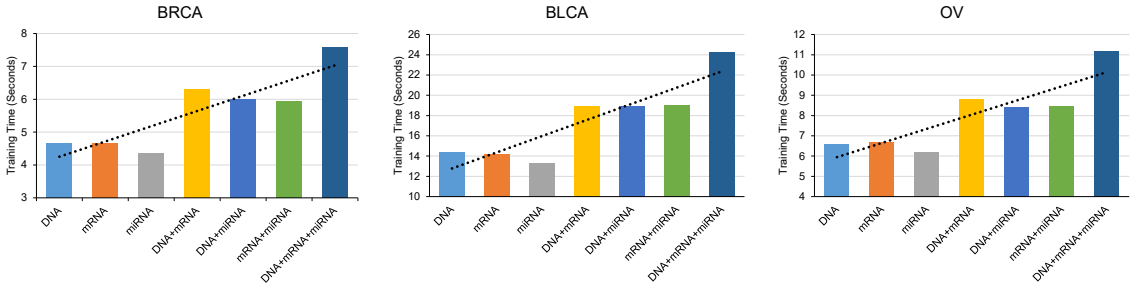

Fig. 5.8 **Average execution time of `MAGNET` over five runs across different omics modality combinations.**

ensuring that all modality combinations have the same number of patients, which allows a fair comparison. We report the average execution time over five runs for different combinations of modalities, and the results are shown in Figure 5.8. We observe that, across all datasets, the runtime increases approximately linearly as more modalities are added. This indicates that the computational cost of `MAGNET` scales linearly with the number of input modalities.

5.5 Summary

In this chapter, we proposed `MAGNET`, a novel method for direct prediction using partial omics modalities, without the need for imputation or patient exclusion, for cancer classification. `MAGNET` introduced a patient-modality multi-head attention mechanism to fuse patient representations by learning the importance of each modality for a patient while adaptively masking out missing modalities. It also constructed a patient interaction graph using the input data with missing modalities for connectivity and the learned embeddings as node

features. To preserve patient-level structure during fusion, MAGNET further incorporated a KL divergence-based loss that aligned similarity distributions before and after fusion. Key strengths of MAGNET include its linear complexity and extensibility with new modalities, as well as its robustness to diverse missing-modality patterns in both training and test data. Experimental results on three multiomics datasets with real-world missingness demonstrated its superiority over baseline methods. Furthermore, quantitative analysis of the learned representations validated its effectiveness.

This chapter focused on biological datasets, specifically multiomics data for cancer classification, allowing targeted evaluation in a clinically relevant context. Future work will explore the integration of fundamentally different modalities, such as imaging and text, to support broader applicability. Although our experiments were limited to cancer-related applications, the proposed method is general and has the potential to be applied to non-biomedical domains.

Chapter 6

Multimodal Feature Selection with Missing Modalities: A Study of Diet and Metformin Effects on the Mouse Brain

In Chapter 5, we introduced a multimodal learning method to tackle missing modalities. However, missing modalities can also be mitigated during feature selection preprocessing before the learning stage. As we described multimodal feature selection methods in Chapters 3 and 4, we assumed complete omics modalities in those settings. Therefore, in this chapter, we address **Research Question 4**, as outlined in Section 1.2, and extend our multimodal feature selection strategy by handling of missing modalities into its process. We propose `MMAgentOmics`, **M**issing-modality-aware multimodal feature selection using a **M**ulti-**agent** system for **omics** integration. In this method, we consider missingness in the computation of static edge weights for within- and cross-modality interactions where missingness can occur. We apply `MMAgentOmics` to a multiomics dataset collected from mice with diet-induced obesity, with and without metformin treatment. This case study is motivated by the fact that chronic obesity is a major global health concern and is associated with an increased risk of diabetes. Metformin is the most widely prescribed drug for the treatment of diabetes. However, the effects of chronic obesity and metformin treatment on

the brain remain poorly understood. In this context, our analysis examines molecular changes associated with diet-induced obesity and metformin treatment.

6.1 Introduction

Metabolic dysfunction associated with obesity is increasingly recognized as a contributor to altered brain physiology and cognitive decline (Schmitt and Gaspar, 2023; Kullmann et al., 2016). High-fat diet (HFD) exposure disrupts metabolic signalling, induces neuroinflammation, and interferes with neuronal function and synaptic regulation (González Olmo et al., 2023; Cavaliere et al., 2019). Metformin, a widely used metabolic therapy, has been reported to modulate systemic energy metabolism and has emerging evidence for brain-related effects (Foretz et al., 2014; Sood et al., 2024). However, the molecular pathways through which diet-induced metabolic stress and metformin intervention influence the brain remain poorly characterized (Li et al., 2022b; Dionysopoulou et al., 2021).

Single-omics analyses often fail to capture the full spectrum of biological alterations underlying complex metabolic and neurological processes. In particular, transcriptional changes alone may not fully reflect metabolic pathway activity, and lipid-mediated signalling plays a key role in neuroinflammation, energy homeostasis, and neuronal function (Bazinet and Layé, 2014). Integrating transcriptomics and lipidomics data therefore provides a more comprehensive molecular view of diet- and drug-related effects in the brain.

A key challenge in this setting is identifying the most informative molecular features from high-dimensional omics data. The problem becomes even more difficult when all data from some modalities are missing for some samples, leading to a missing-modality setting. While several studies have addressed missing modalities during model training, incorporating modality missingness directly into the feature selection process remains much less explored.

In this chapter, we introduce `MMAgentOmics`, a multimodal feature selection framework based on a multi-agent system that operates effectively in the presence of missing modalities. Unlike existing methods, `MMAgentOmics` explicitly handles partial omics modalities by enabling agents to explore a feature space defined by both within- and cross-omics interactions

while accounting for modality missingness during state transitions. Moreover, instead of early fusion, the method evaluates candidate feature subsets using a late fusion strategy, where predictions are aggregated only across the modalities available for each sample.

We assess the performance of `MMAgentOmics` on a multiomics dataset collected from mice with diet-induced obesity, with and without metformin treatment. This work is motivated by two biological questions: (i) whether diet or metformin directly influence the brain, and (ii) how chronic obesity or metformin exposure alters brain molecular signatures.

6.2 Contributions

Our main contributions can be summarized as follows:

- We propose `MMAgentOmics`, an extension of the multimodal feature selection module of `HeteroGATomics`, which can handle missing modalities in its process. It is incorporated into the computation of static edge weights, where missing modalities affect the results.
- We apply `MMAgentOmics` to a real multiomics dataset to study how diet and metformin affect brain molecular profiles in mice. The results show that the features selected by the model align with known biological mechanisms: metformin-associated signatures are consistent with AMPK-mediated energy regulation, while HFD signatures correspond to pathways related to insulin resistance and lipid metabolism.

6.3 Methodology

`MMAgentOmics` is an MAS-based architecture designed for feature selection, in which agents collaboratively explore a network representation of multiomics features to identify the most informative ones in the presence of missing modalities. In the following, we first describe the multiomics feature network representation used as the search space and then explain how the components of the `MMAgentOmics` algorithm differ from those of the multimodal feature selection module in `HeteroGATomics`.

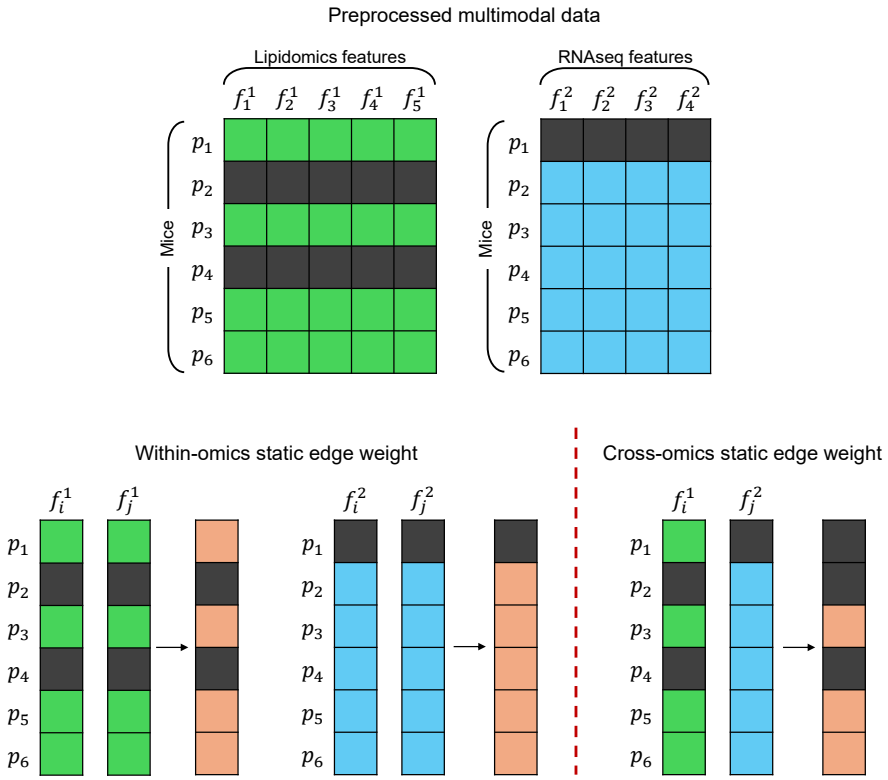

Fig. 6.1 **Static edge weight computation for feature relationships**. Within-omics static edge weights are computed between features of the same omics modality using all available mice. Cross-omics static edge weights are computed between features from different omics modalities using only matched mice. A gray row within a modality indicates a missing modality for that mouse. Gray cells are excluded from the calculation.

6.3.1 Multiomics Feature Network Representation

We consider the same multiomics feature network described in Section 4.3.1. In that section, we assumed that all omics modalities were available for each sample. Here, we modify the definition of the static edge weight to account for missing modalities. Specifically, when calculating the static edge weight between pairs of features within an omics modality, we exclude mice for which that modality is missing. When computing cross-modality edge weights between two features from different omics modalities, we only consider mice that have data available for both modalities. Figure 6.1 provides an example of the computation of both within-omics and cross-omics static edge weights.

6.3.2 The `MMAgentOmics` Framework

`MMAgentOmics` follows the multimodal feature selection procedure described in the framework of `HeteroGATomics` (Section 4.3.1). The overall algorithmic structure remains the same, including the state transition rules, dynamic value updating rules, and omics importance updating rule, which are adopted without modification. The only difference is the fitness function, where $\textsc{Quality}(\cdot)$ quantifies the classifier's performance on the selected feature subsets using a late fusion strategy.

For each modality, we first reduce the training data to the features selected for that modality. We then train and evaluate a classifier independently on each modality using K-fold cross-validation on the training data to obtain class-probability predictions for each mouse (no test data are used at this stage). Because some mice may have missing modalities, predictions are aggregated across the available ones by averaging their probability vectors. The final label for each mouse corresponds to the class with the highest averaged probability.

6.4 Experiments

6.4.1 Data Collection

Data were provided by Dr. Edward C. Harding and Dr. Florian T. Merkle, University of Cambridge. In vivo work was performed under a Home Office licence and approved by the institutional AWERB committee. Briefly, data were collected as follows: Forty male C57BL/6J mice were purchased at 7 weeks old and housed in individually ventilated cages, five per group. They were acclimatized to the facility for 1 week and then habituated to a composition-matched control diet of 10 kcal% fat (D12450J – Research Diets) for a further 1 week before randomization to treatment factors. For the diet factor, mice were randomized to either remain on control diet or switch to a composition-matched 60 kcal% high-fat diet (D12492 – Research Diets). One week later, mice in both control diet and high-fat diet groups were further randomized to treatment arms of vehicle (water) or drug (metformin) for 20 weeks before cardiac blood withdrawal under terminal anaesthesia, concluding with schedule

one culling and tissue collection. Hippocampal samples were freshly micro-dissected and frozen on dry ice. Blood was collected in lithium-heparin-coated tubes and centrifuged at 800 $\times$ g to yield plasma, which was frozen on dry ice. Each hippocampal sample from the same hemisphere was homogenized in a bead homogenizer in 200 μl cold PBS and then split in half for lipidomics and RNAseq using standard protocols at the Institute of Metabolic Science (IMS), University of Cambridge, Lipidomics and Genomics Cores, respectively. Plasma was processed for metabolites and cytokines at the IMS Core Biochemical Assay Laboratory.

6.4.2 Data Preprocessing

Before analysis, we perform four preprocessing steps on the two omics modalities: RNAseq and lipidomics. It is important to note that, unlike the TCGA-based datasets used in previous chapters, this dataset is smaller in terms of sample size and involves different data modalities, which require a different preprocessing strategy and threshold choices to reflect its specific characteristics. First, features in RNAseq with more than 50% missing values are removed, while the lipidomics data contain no missing values. The remaining missing values in RNAseq are imputed using scikit-learn's multivariate feature imputation (Pedregosa et al., 2011) with 1000 features used for imputation and a maximum of 5 iterations. Compared to mean imputation used in Chapters 5, multivariate imputation is adopted here to better capture feature dependencies in RNAseq data in smaller number of samples. Second, features with more than 50% zero values are removed across all modalities, as highly sparse features provide limited discriminatory power and may introduce noise. Third, all features are normalized using min-max scaling (You et al., 2020) to the range [0,1] to ensure comparability across features and to meet the requirements of the `MMAgentOmics` model. Fourth, feature selection is performed based on variance thresholds, retaining those with the highest variance, set to 0.075 for RNAseq and 0.04 for lipidomics. These modality-specific thresholds reflect differences in data distributions and are chosen to balance feature reduction and information preservation. Table 6.1 presents the characteristics of the multiomics dataset following each stage of preprocessing.

Table 6.1 Multiomics data characteristics at each preprocessing stage in the `MMAgentOmics` experiments.

Omics	#Samples	#Original features	#Features after missing value removal	#Features after zero value removal	#Features after variance filtering
RNAseq	24	34,732	34,391	25,146	6,870
Lipidomics	40	1,073	1,073	717	666

6.4.3 Baselines

We compare `MMAgentOmics` with three feature selection methods: mutual information (MI) (Theodoridis and Koutroumbas, 2008), recursive feature elimination (RFE) (Guyon et al., 2002), and minimal-redundancy-maximal-relevance (mRMR) (Peng et al., 2005). To evaluate MI, RFE, and mRMR, we employ the same late fusion strategy described in Section 6.3.2, except that classifiers are trained on the training data and evaluated on the test data. Feature selection is performed independently for each modality using the training set, and both the training and test sets are reduced according to the selected features.

6.4.4 Evaluation Metrics

We consider three tasks in this study. The first task examines the effect of diet, comparing high-fat diet (HFD) mice with control diet (CD) mice. The second task focuses on the effect of the drug, comparing metformin and vehicle groups. The final task is a multi-class classification problem that combines both diet and drug conditions for a more comprehensive analysis. This results in four classes: CD + metformin (CM), CD + vehicle (CV), HFD + metformin (HM), and HFD + vehicle (HV).

The evaluation is performed using accuracy and F1 score for binary classification tasks, and accuracy and weighted F1 score for the multi-class classification task. We report the mean and standard deviation of these metrics across all experiments.

6.4.5 Evaluation Strategy

We evaluate the performance of feature selection methods on the preprocessed dataset, which contains real-world missing modalities. The dataset is first partitioned into matched mice

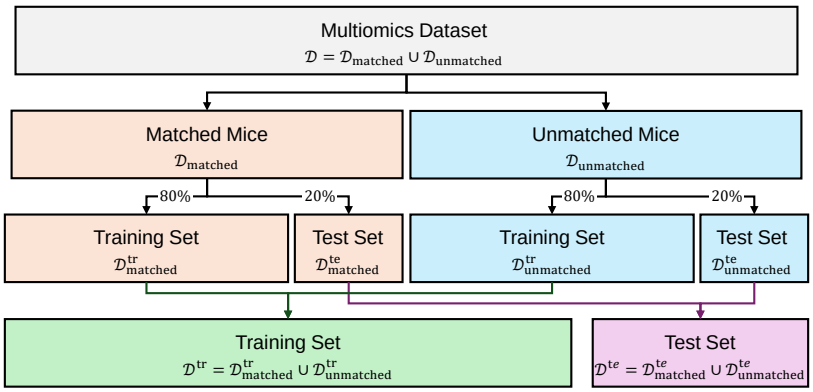

Fig. 6.2 **Data splits for feature selection methods with missing modalities.** The input multiomics data is divided into a matched group, consisting of mice with available data for all omics modalities, and an unmatched group, consisting of mice with missing modalities. Each group is then split into an 80% training set and a 20% test set. Finally, matched and unmatched mice within each split are combined to form the final training and test sets.

$\mathcal{D}_{\text{matched}}$, which have both RNAseq and lipidomics modalities available, and unmatched mice $\mathcal{D}_{\text{unmatched}}$, which lack one of these modalities. Each group is then split separately into a training set (80%) and a test set (20%) using a stratified sampling strategy. The matched and unmatched mice within each split are then combined to form the final training set $\mathcal{D}^{\text{tr}}$ and test set $\mathcal{D}^{\text{te}}$. To ensure robust performance estimation, this procedure is repeated 10 times, each with a different random partition of the training and test sets. Figure 6.2 illustrates the data splitting process.

6.4.6 Implementation Details

`MMAgentOmics` has been developed using Python 3.10, incorporating essential functionalities from scikit-learn 1.3.0 (Pedregosa et al., 2011), NumPy 1.26.0 (Harris et al., 2020), pandas 2.1.1 (The pandas development team, 2023), and SciPy 1.11.3 (Virtanen et al., 2020). For `MMAgentOmics`, the following parameters are used: the maximum number of iterations $T = 50$, the number of agents per omics modality $N_A = 10$, node decay coefficient $\rho_{\mathcal{V}} = 0.1$ in Equation (3.5), edge decay coefficient $\rho_{\mathcal{E}} = 0.1$ in Equation (4.5), omics importance decay coefficient $\gamma = 0.1$ in Equation (3.8), the initial dynamic node score for each node $\tau_u^i(0) = 0.2$, the initial dynamic edge weight for each edge $\tau_{uv}^i(0) = 0.2$, and the state

transition rule control parameter $q_0 = 0.8$. We set $K = 5$ in the K-fold cross-validation used within the late fusion strategy for computing the fitness function.

We utilize the scikit-feature package (Li et al., 2018) for implementing mRMR, and scikit-learn for implementing MI and RFE.

6.4.7 Feature Selection Performance Comparison

Figure 6.3 shows the performance of the feature selection methods using the RF classifier with different numbers of selected features (from 10 to 100) across three tasks. In the diet factor classification task (Figure 6.3a), `MMAgentOmics` outperforms all baselines in both accuracy and F1 score for most feature sizes. In the drug factor classification task (Figure 6.3b), `MMAgentOmics` continues to excel in identifying informative features, surpassing the best performance of MI by 5.5%, RFE by 3.3%, and mRMR by 7.7% in accuracy, and the best performance of MI by 7.0%, RFE by 0.7%, and mRMR by 7.1% in F1 score. In the combined diet and drug factor classification task (Figure 6.3c), although MI performs better than `MMAgentOmics` at certain feature sizes, `MMAgentOmics` remains the top overall performer, achieving the highest classification accuracy and weighted F1 score compared with the baselines. Notably, as the number of selected features increases, `MMAgentOmics` tends to show improved performance, whereas MI exhibits less consistent behaviour. This can be attributed to the fact that `MMAgentOmics` is designed to capture complementary interactions among features, allowing it to benefit from a larger feature subset, while MI evaluates features independently and may include redundant or less informative features as the feature set grows. Additionally, we observe that the effectiveness of a feature selection method strongly depends on the classification task. For example, MI performs well in the combined diet and drug factor classification task but is among the weakest in the drug factor classification task, which is the most challenging classification problem among the three. These results highlight `MMAgentOmics`'s ability to identify discriminative features effectively across different tasks.

To further assess the discriminative ability of the features selected by `MMAgentOmics`, we conduct a receiver operating characteristic (ROC) analysis for the diet and drug factor

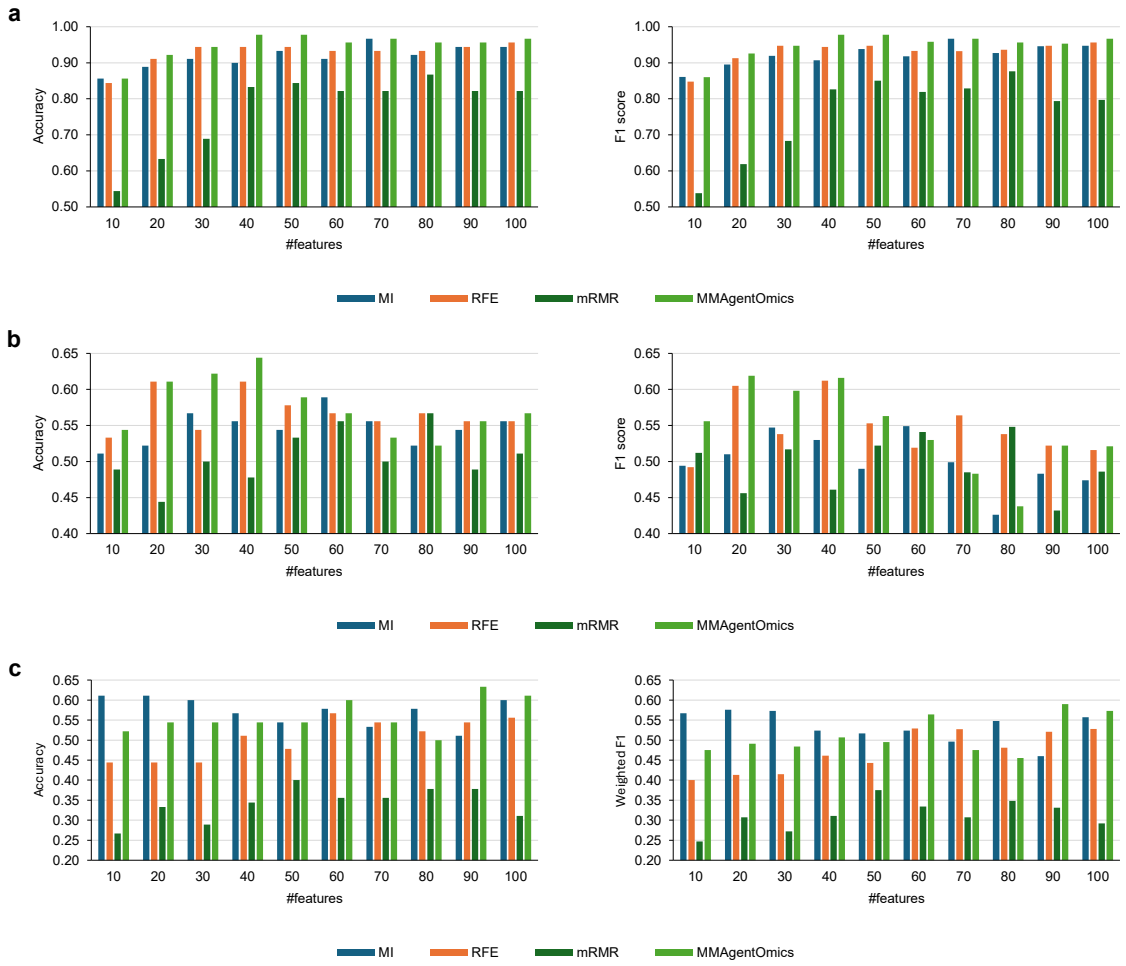

Fig. 6.3 **Performance comparison of feature selection methods using the RF classifier.** **a**, Diet factor classification. **b**, Drug factor classification. **c**, Combined diet and drug factor classification.

classification tasks. Figure 6.4 presents the ROC analysis of classification performance across three classifiers using 60 selected features identified by `MMAgentOmics` over 10 independent runs. The ROC curve plots sensitivity against 1 – specificity, providing an overview of how well each classifier separates classes at different decision thresholds. The diet factor task shows near-perfect discrimination, with a mean AUC of 0.99 across all three classifiers, whereas the drug factor task shows lower and more variable performance, with mean AUC values of 0.65, 0.65, and 0.61 for RF, *k*NN, and SVM, respectively. These results indicate that the diet factor classification is substantially easier and more reliable, while the drug

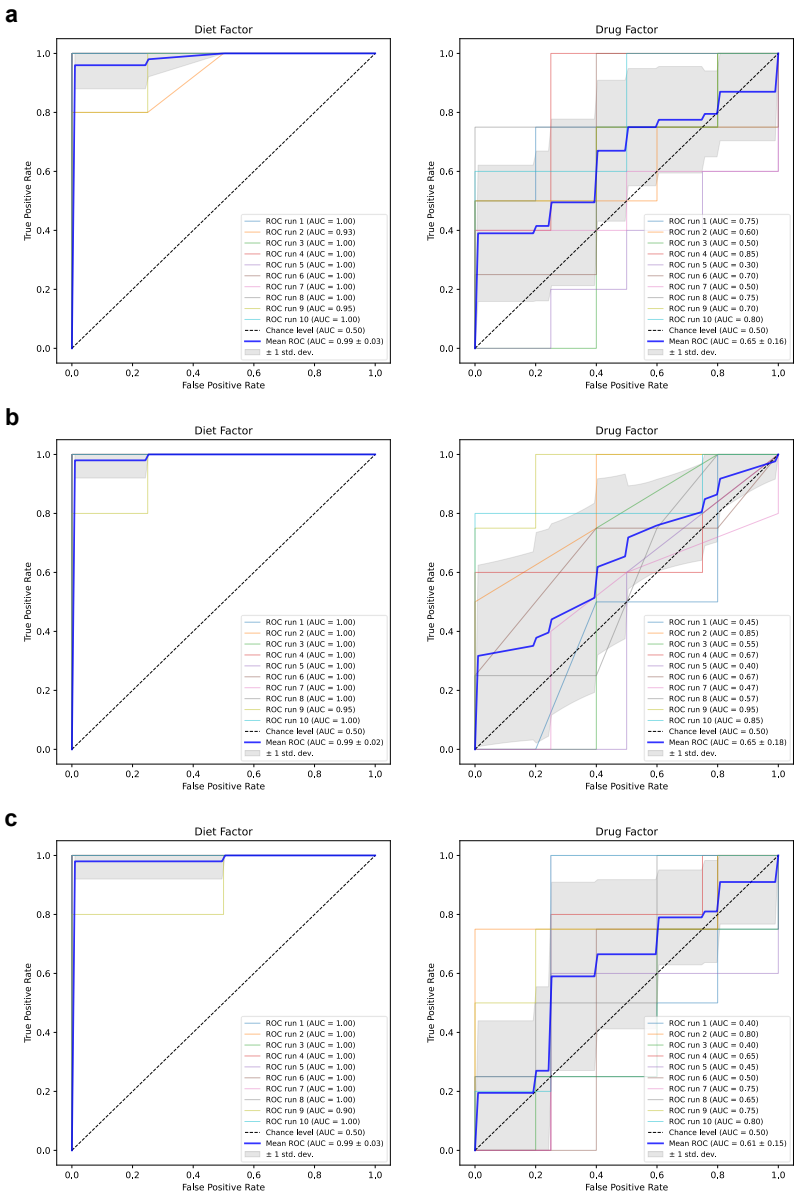

Fig. 6.4 **ROC analysis of the classification performance for the diet and drug factors using 60 selected features identified by `MMAgentOmics` across 10 independent runs. a**, RF classifier results. **b**, *k*NN classifier results. **c**, SVM classifier results. The dashed diagonal curve represents chance-level performance, corresponding to a classifier with no ability to discriminate between classes.

factor task remains more challenging. This observation is consistent with previous findings. The ROC analysis therefore supports placing greater interpretive weight on the diet factor results, as the model achieves a higher level of confidence in that setting.

6.4.8 UMAP Visualization of Feature Representations

In order to evaluate the effectiveness of the features selected by `MMAgentOmics` as a multimodal feature selection method, we visualize the UMAP representations of samples for three tasks under four conditions: original features from each individual omics modality (RNAseq and lipidomics), concatenation of the all features from both modalities, and concatenation of the 60 features selected across both modalities using `MMAgentOmics`. Figure 6.5 shows that the original features form a highly entangled structure with dispersed sample distributions, making class separation challenging. Even after concatenating features from both modalities, the integrated feature space still does not achieve clear separation. In contrast, `MMAgentOmics` achieves clear class separation across all tasks, as evidenced by higher intra-cluster compactness and greater inter-cluster separation, highlighting the effectiveness of multimodal feature selection in capturing meaningful biological patterns. This visualization provides qualitative evidence that `MMAgentOmics` effectively captures task-relevant feature structures that enhance class separability.

6.4.9 Analysis of the Top-ranked Features

To gain insights into feature importance and patterns identified by `MMAgentOmics`, and to facilitate biological interpretation, we analyze the top selected features. We generate K independent random splits of the dataset and run the feature selection algorithm on the training set of each split. The resulting K feature lists are then aggregated into a final ranked list (Guyon and Elisseeff, 2003; Pes et al., 2017) using frequency-based feature ranking strategy (El-Manzalawy et al., 2018; Zucknick et al., 2008).

Top-ranked feature identification technique

We employ a frequency-based feature ranking technique (El-Manzalawy et al., 2018; Zucknick et al., 2008) to identify the most consistently selected features across multiple runs. This technique first identifies the top-ranked features selected by `MMAgentOmics` in each split. Since the selected feature lists may vary across runs, the strategy ranks features by their

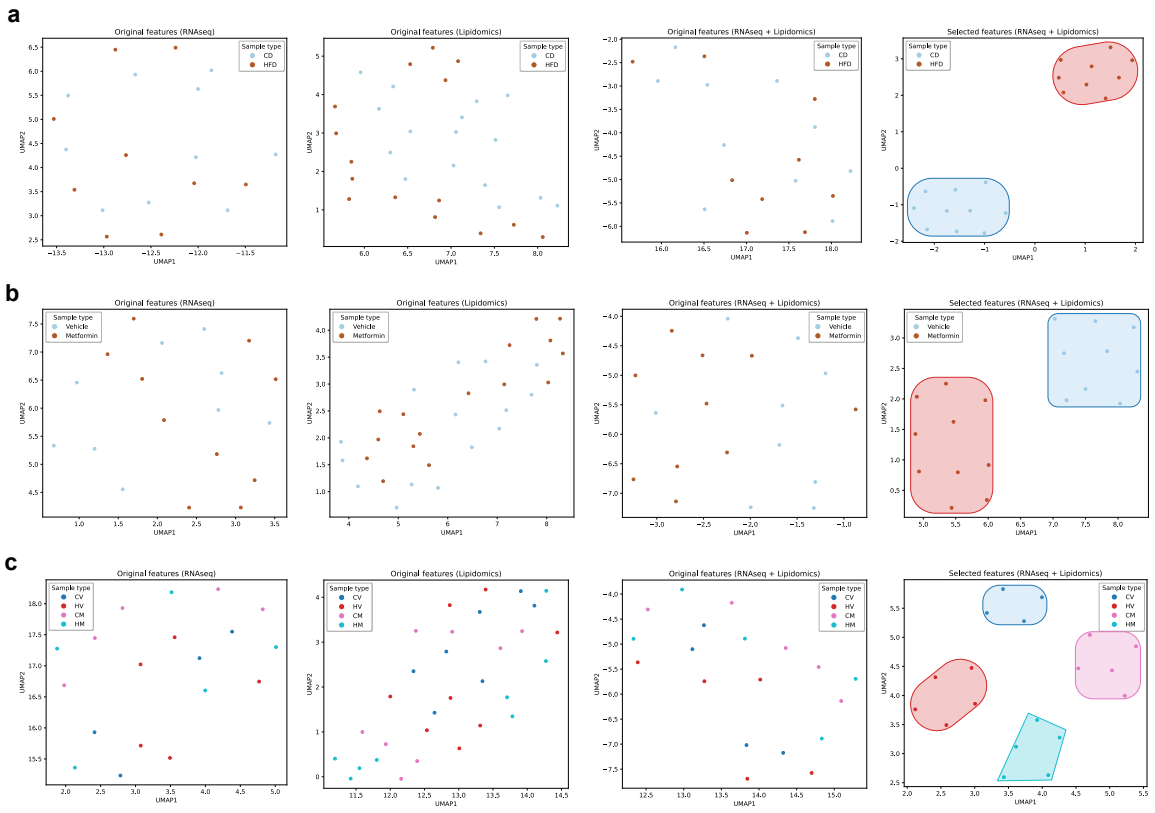

Fig. 6.5 **UMAP visualization of training samples based on original and selected features, colored by sample type.** **a**, Diet factor representation. **b**, Drug factor representation. **c**, Combined diet and drug factor representation. From left to right: original RNAseq features, original lipidomics features, concatenation of original RNAseq and lipidomics features, and concatenation of 60 selected RNAseq and lipidomics features identified by `MMAgentOmics`.

selection frequency—how often they are chosen across all splits. A feature's importance is therefore reflected by its frequency, where features selected more consistently are considered more relevant. The score of a feature f_u is computed as follows:

$$\text{FREQUENCY}(f_u) = \frac{1}{K} \sum_{k=1}^{K} I(f_u \in \mathcal{S}_{\mathcal{V}}(k)) \tag{6.1}$$

where K is the total number of runs, $\mathcal{S}_{\mathcal{V}}(k)$ is the set of selected features in run k, and $I(f_u \in \mathcal{S}_{\mathcal{V}}(k))$ is an indicator function that equals 1 if feature f_u is selected in run k, and 0 otherwise.

Results and analysis

Table A.2 in Appendix A presents the results of the frequency-based ranking strategy, highlighting the top-ranked features identified across 10 runs. We consider the top 60 features selected by `MMAgentOmics` in each run and, in the end, retain only those with high confidence of being expressed coding genes.

The top-ranked coding genes highlight biologically meaningful patterns across the diet and drug tasks. Diet-associated genes such as Ifitm7, Hmga2, and Adcy4 are linked to immune activation (Friedlová et al., 2022), neuronal signaling (Devasani and Yao, 2022), and metabolic regulation (Xi et al., 2016), consistent with known effects of chronic HFD–induced obesity on inflammation and brain metabolism (Evans et al., 2024). Metformin-associated genes, including Banp and Sh3bp5, relate to mitochondrial activity, cell-cycle control (Babu et al., 2022), and intracellular signaling pathways (Mansouri et al., 2018), aligning with metformin's reported effects on cellular energy metabolism and AMPK-mediated mechanisms (Goel et al., 2022; Ma et al., 2022b). Genes shared in the combined task (Cpt1b, Acot3, Slc2a2) are predominantly involved in fatty-acid oxidation (Wang et al., 2021a) and glucose transport (Morrice et al., 2023), suggesting integrated regulation of energy and nutrient metabolism when both diet and drug factors are considered.

To evaluate the biological relevance of the features selected by `MMAgentOmics`, we perform pathway enrichment analysis on the top-ranked genes for each task. This analysis examines whether the identified molecular signatures align with known pathways related to HFD, metformin, and their combined effects. Pathway enrichment is conducted using ShinyGO v0.85 (Ge et al., 2020) with KEGG as the reference database, and enriched pathways are ranked by fold enrichment. Results are shown in Figure A.4 in Appendix A.

Diet factor enrichment. For the diet task (Figure A.4a), the most enriched pathways are primarily associated with metabolic and hormone-regulated signaling processes, including insulin secretion, GABAergic and cholinergic synapse signaling, circadian regulation, and lipid metabolism. These enrichment patterns align with established effects of HFD exposure, which has been reported to disrupt central insulin signaling, alter hypothalamic nutrient-

sensing circuitry, and influence synaptic and circadian processes (Cai and Liu, 2011; Timper and Brüning, 2017; Thaler et al., 2012). These findings suggest that the model may capture brain-relevant molecular signatures associated with HFD-induced metabolic and neuronal rewiring.

Drug factor enrichment. For the drug task (Figure A.4b), fewer pathways reach significance, indicating that metformin-related effects are less clearly captured in this dataset compared to diet-related changes. The top enriched pathways include glycosaminoglycan biosynthesis, unsaturated fatty-acid biosynthesis, folate metabolism, and steroidogenesis. These patterns are consistent with reported effects of metformin on metabolic signaling and mitochondrial function, including its role in modulating lipid handling and one-carbon metabolic processes through AMPK-related mechanisms (Foretz et al., 2019; Rena et al., 2017). Overall, while some pathways align with known metformin biology, the metformin-related effects in this dataset remain unclear.

Combined diet and drug enrichment. For the combined task (Figure A.4c), the enriched pathways span metabolic and endocrine processes, including fatty-acid metabolism, adipocytokine and insulin-related signaling, amino-acid biosynthesis, and glucagon signaling. These results show that pathways associated with both diet and drug conditions appear in the combined setting, without implying direct interaction between the two. Such patterns are broadly consistent with reports that diet-induced metabolic stress and metformin-mediated energy regulation influence overlapping nutrient-sensing and metabolic pathways (Newgard et al., 2009; Pernicova and Korbonits, 2014). This enrichment suggests that the model may capture shared molecular signatures associated with chronic HFD exposure and metformin treatment in this dataset.

Taken together, these results indicate that `MMAgentOmics` not only distinguishes experimental groups, but also identifies feature patterns enriched in biologically meaningful pathways consistent with reported diet- and metformin-related metabolic effects. This suggests that the model may capture relevant molecular structure rather than relying solely on statistical separation.

6.5 Summary

In this chapter, we presented `MMAgentOmics`, a multimodal feature selection framework designed to operate under missing-modality conditions. The method uses a multi-agent system with a late fusion evaluation strategy to identify informative features when only partial omics data are available, incorporating both within- and cross-omics interactions during state transitions. Across the diet and drug classification tasks, `MMAgentOmics` outperforms baseline feature selection approaches across most feature sizes and demonstrates improved robustness to missing modalities. In the combined diet and drug classification task, its performance is comparable to MI, yet `MMAgentOmics` remains the overall best performer, achieving the highest accuracy and weighted F1 score. Applied to a real multiomics mouse dataset, `MMAgentOmics` identifies molecular signatures consistent with known biological effects of dietary and metformin interventions on the brain. The selected features and enriched pathways reflect AMPK-related metabolic regulation under metformin treatment and insulin- and lipid-related alterations under HFD exposure, supporting biologically coherent interpretation of model outputs. Future work will involve applying `MMAgentOmics` to larger multiomics datasets to further assess its scalability and translational potential.

Chapter 7

Conclusion and Future Work

In this thesis, we proposed four methods for multimodal learning with graphs to improve cancer classification using multiomics data. Overall, our work explores three main challenges in the general workflow of multimodal learning with graphs (Figure 1.1) and addresses four research questions (Section 1.2) formulated to overcome these challenges. We summarize our contributions in addressing these research questions in Section 7.1 and outline promising directions for future research in Section 7.2.

7.1 Conclusion

This thesis aimed to address four key research questions in multimodal learning with graphs.

Research Question 1: *How can we develop a multimodal feature selection method that reduces omics feature dimensionality while accounting for both intra- and inter-omics relationships?*

With the availability of high-dimensional omics data, feature selection plays an important role in enabling models to learn from informative features. However, existing methods are often applied independently to each omics modality, thereby overlooking potential relationships across modalities. In Chapter 3, we addressed this limitation by introducing `MAgentOmics`, a multimodal feature selection method based on a multi-agent system. We constructed a multiomics feature network, a graph-based framework in which features from

different modalities are represented as nodes and their relationships as edges. This representation enables the multi-agent system to explore the feature space while accounting for both intra- and inter-omics relationships, ultimately leveraging complementary information across modalities to identify more informative and biologically meaningful features.

Research Question 2: *How can we construct heterogeneous graphs to learn holistic graph representations that capture diverse structures in multiomics data?*

While homogeneous graphs can lose structural information and the diverse nature of multiomics data, heterogeneous graphs model multiple types of nodes and edges, thereby better reflecting the complexity of multiomics relationships. However, the construction of heterogeneous graphs in existing works typically relies on pre-existing knowledge graphs, which are not always available or may be costly to build. In Chapter 4, we introduced `HeteroGATomics`, a dual-view framework that automatically constructs heterogeneous graphs by leveraging auxiliary information inherent in multiomics data. The heterogeneous graph is formed by combining a patient similarity network that captures interactions among patients and a feature similarity network that captures relationships among features. Predictions are then generated on the modality-specific heterogeneous graph using a GNN, and late fusion integrates these predictions in a supervised manner. This framework facilitates cross-modality interactions at both the feature and label levels.

Research Question 3: *How can we integrate multiomics data in the presence of different missing-modality patterns, enabling predictions directly from partial modalities?*

Graph-based multimodal methods often assume that all omics modalities are available for each patient. However, missing modalities are common in real-world multimodal biomedical settings. While prior approaches address this challenge, many cannot handle different missing-modality patterns across training and test data or scale with the number of modalities. In Chapter 5, we presented `MAGNET`, a framework for direct prediction with partial modalities. `MAGNET` introduces a patient-modality multi-head attention mechanism to fuse modality embeddings based on their importance and missingness. Its complexity scales linearly with the number of modalities, allowing adaptation to missing-pattern variability.

We also incorporated modality missingness to construct a patient interaction graph, with fused multimodal embeddings as node features and connectivity determined by modality availability. This aligns with clinical reasoning, where patients who share characteristics in available modalities are likely to show similarities in missing ones.

Research Question 4: *How can we develop a multimodal feature selection method that explicitly incorporates missing modalities into the selection process?*

With the presence of missing modalities, feature selection is still often performed only on complete data, thereby excluding cases with missing modalities. In Chapter 6, we developed `MMAgentOmics` to enable multimodal feature selection in the presence of missing modalities. This method extends our multimodal feature selection strategy based on a multi-agent system by adapting the selection of features within and across modalities to account for missing modalities. We validated this algorithm on a real multiomics case study investigating the effects of diet-induced obesity and metformin treatment on molecular changes in mice. The selected features were consistent with known biological mechanisms, demonstrating the biological relevance of our method.

7.2 Future Work

The development of new architectures for multimodal learning with graphs remains a broad and evolving research area. In this thesis, we introduced novel methods and findings that open several promising directions for further investigation. In this section, we outline potential avenues for future work.

7.2.1 Deep Learning-Driven Multimodal Feature Selection

In Chapters 3 and 6, we introduced multimodal feature selection frameworks that extend the concept of MAS to multimodal settings. This formulation provides a foundation for adapting other feature selection algorithms originally designed for unimodal data to multimodal learning scenarios. While the multi-agent paradigm can be parallelized to mitigate

computational cost, adding more modalities with high dimensional features increases the number of nodes and edges, which may still lead to scalability challenges. A potential future direction is to explore deep learning for multimodal feature selection, where neural networks assist agent decisions to efficiently navigate large feature graphs. Inspired by the progression from classical reinforcement learning based feature selection (Liu et al., 2021) to deep reinforcement learning (Potharlanka and M, 2024; Liu et al., 2023), such an approach can reduce search cost and make the process more scalable for large multimodal graphs.

7.2.2 Constructing a Unified Multi-Relational Graph Framework

In Chapter 4, we constructed modality-specific heterogeneous graphs and applied GNNs to learn patient representations from each modality, followed by late fusion to obtain final predictions. While this approach demonstrated the advantages of graph-based modeling, it treats modalities independently and only combines information at the prediction stage. A promising future direction is to develop a single unified heterogeneous graph that jointly represents patients and multiomics features within one structure. Such a graph would include multiple node types (e.g., patients, genes, proteins) and edge types (e.g., feature–feature associations within a modality, cross-omics feature interactions, patient–patient similarity, and feature–patient relationships). By capturing both intra- and cross-modal biological dependencies within one graph, a unified model can propagate richer information between molecular features and patients, enabling a more holistic and biologically grounded representation. A single GNN operating on this unified graph could therefore learn more expressive and integrated patient embeddings, potentially improving predictive performance and interpretability compared to late fusion schemes.

7.2.3 Modeling Missing Values Within Heterogeneous Graphs

Although the primary focus of this thesis was handling missing modalities, addressing missing values within each modality is an important extension. The framework introduced in Chapter 4 provides a foundation for this direction, as it naturally supports missing values in

multiomics data without requiring imputation or discarding patients. In this thesis, however, missing values were removed during preprocessing, and the feature–patient relationships in the graph assumed that every feature was observed for every patient. A valuable future direction is to extend this framework to operate directly on raw data containing missing values. In this setting, edges between a feature node and a patient node would be removed whenever that feature is missing for that patient. This yields modality-specific heterogeneous graphs in which only edges corresponding to observed values remain. A GNN operating on this structure could then learn patient representations while inherently modeling missing values, eliminating the need for imputation and preserving the original data structure. Such an approach would enable more principled learning from sparse multiomics datasets, a common scenario in biomedical applications.

7.2.4 Integrating Additional Data Modalities

Multiomics data include molecular, imaging, and phenotypic information. This thesis primarily focused on molecular omics data, enabling controlled evaluation of the proposed methods within a cancer classification context. However, human diseases arise from complex interactions across biological, spatial, and clinical dimensions. Incorporating fundamentally different data types, such as histopathology or radiology images, electronic health records, and clinical notes, would provide a more comprehensive view of disease processes and strengthen clinical relevance (Acosta et al., 2022). Future work can therefore extend the current framework to integrate modalities beyond molecular profiles, enabling multimodal reasoning across structured omics data, unstructured clinical text, and medical imaging. Such extensions would broaden the applicability of the proposed models and support their use in translational and real-world settings beyond oncology, including general biomedical and potentially non-biomedical domains.

7.2.5 Multimodal Fusion with Modality Selection

As an increasing number of diverse data modalities become available, an important question in multimodal analysis is whether incorporating additional modalities truly improves biological insight and predictive performance. Different modalities have distinct structures and information content, and combining them randomly can introduce redundancy or even degrade model performance. This motivates the development of approaches for modality selection, determining when and how each modality should contribute to the learning process. Future work will explore adaptive multimodal fusion strategies that selectively integrate the most informative modalities and identify optimal modality combinations for specific biological tasks.

References

Acosta, J. N., Falcone, G. J., Rajpurkar, P., and Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine*, 28(9):1773–1784.

Adossa, N., Khan, S., Rytkönen, K. T., and Elo, L. L. (2021). Computational strategies for single-cell multi-omics integration. *Computational and Structural Biotechnology Journal*, 19:2588–2596.

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., and Narasimhan, G. (2016). Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis: supplementary issue: bioinformatics methods and applications for big metagenomics data. *Evolutionary Bioinformatics*, 12:EBO–S36436.

Alkhateeb, A., Tabl, A. A., and Rueda, L. (2021). Deep learning in multi-omics data integration in cancer diagnostic. In *Deep Learning for Biomedical Data Analysis*, pages 255–271. Springer.

Amjad, R. A., Liu, K., and Geiger, B. C. (2022). Understanding neural networks and individual neuron importance via information-ordered cumulative ablation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7842–7852.

Amor, A., Lio, P., Singh, V., Torné, R. V., and Terre, H. A. (2021). Graph representation learning on tissue-specific multi-omics. *arXiv preprint arXiv:2107.11856*.

Aparisi, F., Amado-Labrador, H., Calabuig-Fariñas, S., Torres, S., and Herreros-Pomares, A. (2019). Passenger mutations in cancer evolution. *Cancer Reports and Reviews*, 3(10.15761).

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124.

Arslan, E., Schulz, J., and Rai, K. (2021). Machine learning in epigenomics: insights into cancer biology and medicine. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1876(2):188588.

Ashuach, T., Gabitto, M. I., Koodli, R. V., Saldi, G.-A., Jordan, M. I., and Yosef, N. (2023). MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8):1222–1231.

Austin, P. C., White, I. R., Lee, D. S., and van Buuren, S. (2021). Missing data in clinical research: a tutorial on multiple imputation. *Canadian Journal of Cardiology*, 37(9):1322–1331.

Babu, S., Takeuchi, Y., and Masai, I. (2022). Banp regulates DNA damage response and chromosome segregation during the cell cycle in zebrafish retina. *Elife*, 11:e74611.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: a survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Banks, R. E., Dunn, M. J., Hochstrasser, D. F., Sanchez, J.-C., Blackstock, W., Pappin, D. J., and Selby, P. J. (2000). Proteomics: new perspectives, new biomedical opportunities. *The Lancet*, 356(9243):1749–1756.

Barefoot, M. E., Varghese, R. S., Zhou, Y., Poto, C. D., Ferrarini, A., and Ressom, H. W. (2019). Multi-omic pathway and network analysis to identify biomarkers for hepatocellular carcinoma. In *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1350–1354.

Bayoudh, K., Knani, R., Hamdaoui, F., and Mtibaa, A. (2022). A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970.

Bazinet, R. P. and Layé, S. (2014). Polyunsaturated fatty acids and their metabolites in brain function and disease. *Nature Reviews Neuroscience*, 15(12):771–785.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLOS Computational Biology*, 4(10):e1000173.

Bi, X.-A., Shen, W., Shan, Y., Chen, D., Xu, L., Chen, K., and Liu, Z. (2025). MSAFF: multi-way soft attention fusion framework with the large foundation models for the diagnosis of Alzheimer's disease. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):17541–17555.

Bica, I., Velickovic, P., Xiao, H., and Li, P. (2018). Multi-omics data integration using cross-modal neural networks. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*.

Bilder, R. M., Sabb, F., Cannon, T., London, E., Jentsch, J., Parker, D. S., Poldrack, R., Evans, C., and Freimer, N. (2009). Phenomics: the systematic study of phenotypes on a genome-wide scale. *Neuroscience*, 164(1):30–42.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Bodein, A., Scott-Boyer, M.-P., Perin, O., Lê Cao, K.-A., and Droit, A. (2022). Interpretation of network-based integration from multi-omics longitudinal data. *Nucleic Acids Research*, 50(5):e27–e27.

Bouchard, N. and Daaboul, N. (2025). Lung cancer: targeted therapy in 2025. *Current Oncology*, 32(3).

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brown, J. S., Amend, S. R., Austin, R. H., Gatenby, R. A., Hammarlund, E. U., and Pienta, K. J. (2023). Updating the definition of cancer. *Molecular Cancer Research*, 21(11):1142–1147.

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345.

Cai, D. and Liu, T. (2011). Hypothalamic inflammation: a double-edged sword to nutritional diseases. *Annals of the New York Academy of Sciences*, 1243(1):E1–E39.

Cai, L., Wang, Z., Gao, H., Shen, D., and Ji, S. (2018). Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1158–1166.

Cano-Gamez, E. and Trynka, G. (2020). From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*, 11:424.

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., and Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, 12(1):124.

Carrillo-Perez, F., Morales, J. C., Castillo-Secilla, D., Gevaert, O., Rojas, I., and Herrera, L. J. (2022). Machine-learning-based late fusion on multi-omics and multi-scale data for non-small-cell lung cancer diagnosis. *Journal of Personalized Medicine*, 12(4):601.

Caruso, C. M., Soda, P., and Guarrasi, V. (2025). MARIA: a multimodal transformer model for incomplete healthcare data. *Computers in Biology and Medicine*, 196:110843.

Caruso, D., Polici, M., Zerunian, M., Pucciarelli, F., Guido, G., Polidori, T., Landolfi, F., Nicolai, M., Lucertini, E., Tarallo, M., et al. (2021). Radiomics in oncology, part 1: technical principles and gastrointestinal application in CT and MRI. *Cancers*, 13(11):2522.

Cascio, C. L., McNamara, J. B., Melendez, E. L., Lewis, E. M., Dufault, M. E., Sanai, N., Plaisier, C. L., and Mehta, S. (2021). Nonredundant, isoform-specific roles of HDAC1 in glioma stem cells. *JCI Insight*, 6(17).

Cavaliere, G., Trinchese, G., Penna, E., Cimmino, F., Pirozzi, C., Lama, A., Annunziata, C., Catapano, A., Mattace Raso, G., Meli, R., et al. (2019). High-fat diet induces neuroinflammation and mitochondrial impairment in mice cerebral cortex and synaptic fraction. *Frontiers in Cellular Neuroscience*, 13:509.

Cavill, R., Jennen, D., Kleinjans, J., and Briedé, J. J. (2016). Transcriptomic and metabolomic data integration. *Briefings in Bioinformatics*, 17(5):891–901.

Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., Wang, J., Rudolph, J. E., Yaeger, R., Soumerai, T., Nissan, M. H., et al. (2017). OncoKB: a precision oncology knowledge base. *JCO Precision Oncology*, 2017:PO.17.00011.

Chandak, P., Huang, K., and Zitnik, M. (2023). Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Chen, F., Cai, G., Li, Y., and Ou-Yang, L. (2025). SpaFusion: a multi-level fusion model for clustering spatial multi-omics data. *Information Fusion*, page 103372.

Chen, F., Zhang, Y., Şenbabaoğlu, Y., Ciriello, G., Yang, L., Reznik, E., Shuch, B., Micevic, G., De Velasco, G., Shinbrot, E., et al. (2016). Multilevel genomics-based taxonomy of renal cell carcinoma. *Cell Reports*, 14(10):2476–2489.

Chen, J., Yang, S., Peng, X., Peng, D., and Wang, Z. (2022). Augmented sparse representation for incomplete multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):4058–4071.

Chen, M., Zhang, Y., Kou, X., Li, Y., and Zhang, Y. (2021a). r-GAT: relational graph attention network for multi-relational graphs. *arXiv preprint arXiv:2109.05922*.

Chen, R. J., Lu, M. Y., Weng, W.-H., Chen, T. Y., Williamson, D. F., Manz, T., Shady, M., and Mahmood, F. (2021b). Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4025.

Chen, T. and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, page 785–794.

Chen, Y., Barefoot, M. E., Varghese, R. S., Wang, K., Di Poto, C., and Ressom, H. W. (2020). Integrative analysis to identify race-associated metabolite biomarkers for hepatocellular carcinoma. In *42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5300–5303.

Chierici, M., Bussola, N., Marcolini, A., Francescatto, M., Zandonà, A., Trastulla, L., Agostinelli, C., Jurman, G., and Furlanello, C. (2020). Integrative network fusion: a multi-omics approach in molecular profiling. *Frontiers in Oncology*, 10:1065.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057.

Colland, F., Jacq, X., Trouplin, V., Mougin, C., Groizeleau, C., Hamburger, A., Meil, A., Wojcik, J., Legrain, P., and Gauthier, J.-M. (2004). Functional proteomics mapping of a human signaling pathway. *Genome Research*, 14(7):1324–1332.

Collier, M., Nazabal, A., and Williams, C. (2020). VAEs in the presence of missing data. In *ICML Workshop on the Art of Learning with Missing Values (Artemiss)*.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., Zepeda-Mendoza, M. L., Campos, P. F., Heller, R., and Pereira, R. J. (2016). Next-generation biology: sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, 30:3–13.

Dagogo-Jack, I. and Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2):81–94.

Das, S., Hayden, J., Sullivan, T., and Rieger-Christ, K. (2023). The roles of miRNAs in predicting bladder cancer recurrence and resistance to treatment. *International Journal of Molecular Sciences*, 24(2):964.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

De Meulder, B., Lefaudeux, D., Bansal, A. T., Mazein, A., Chaiboonchoe, A., Ahmed, H., Balaur, I., Saqi, M., Pellet, J., Ballereau, S., et al. (2018). A computational framework for complex disease stratification from multiple large-scale datasets. *BMC Systems Biology*, 12(1):60.

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2019). Multi-study factor analysis. *Biometrics*, 75(1):337–346.

Deng, L., Chao, H., Deng, H., Yu, Z., Zhao, R., Huang, L., Gong, Y., Zhu, Y., Wang, Q., Li, F., et al. (2022). A novel and sensitive DNA methylation marker for the urine-based liquid biopsies to detect bladder cancer. *BMC Cancer*, 22(1):510.

Devasani, K. and Yao, Y. (2022). Expression and functions of adenylyl cyclases in the CNS. *Fluids and Barriers of the CNS*, 19(1):23.

Dias-Audibert, F. L., Navarro, L. C., de Oliveira, D. N., Delafiori, J., Melo, C. F. O. R., Guerreiro, T. M., Rosa, F. T., Petenuci, D. L., Watanabe, M. A. E., Velloso, L. A., et al. (2020). Combining machine learning and metabolomics to identify weight gain biomarkers. *Frontiers in Bioengineering and Biotechnology*, 8:6.

Dibitetto, D., Widmer, C. A., and Rottenberg, S. (2024). PARPi, BRCA, and gaps: controversies and future research. *Trends in Cancer*, 10(9):857–869.

Ding, D. Y., Li, S., Narasimhan, B., and Tibshirani, R. (2022). Cooperative learning for multiview analysis. *Proceedings of the National Academy of Sciences*, 119(38):e2202113119.

Dionysopoulou, S., Charmandari, E., Bargiota, A., Vlahos, N. F., Mastorakos, G., and Valsamakis, G. (2021). The role of hypothalamic inflammation in diet-induced obesity and its association with cognitive and mood disorders. *Nutrients*, 13(2):498.

Dong, X., Lin, L., Zhang, R., Zhao, Y., Christiani, D. C., Wei, Y., and Chen, F. (2019). TOBMI: trans-omics block missing data imputation using a k-nearest neighbor weighted approach. *Bioinformatics*, 35(8):1278–1283.

Dorigo, M. and Blum, C. (2005). Ant colony optimization theory: a survey. *Theoretical Computer Science*, 344(2-3):243–278.

Dorigo, M. and Gambardella, L. M. (1997). Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation*, 1(1):53–66.

Dorigo, M. and Stützle, T. (2018). Ant colony optimization: overview and recent advances. *Handbook of Metaheuristics*, pages 311–351.

Dorri, A., Kanhere, S. S., and Jurdak, R. (2018). Multi-agent systems: A survey. *IEEE Access*, 6:28573–28593.

Eddy, S. R. (2001). Non–coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–929.

Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M., and Zitnik, M. (2023). Multimodal learning with graphs. *Nature Machine Intelligence*, 5(4):340–350.

El-Manzalawy, Y., Hsieh, T.-Y., Shivakumar, M., Kim, D., and Honavar, V. (2018). Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data. *BMC Medical Genomics*, 11(3):19–31.

Evans, A. K., Saw, N. L., Woods, C. E., Vidano, L. M., Blumenfeld, S. E., Lam, R. K., Chu, E. K., Reading, C., and Shamloo, M. (2024). Impact of high-fat diet on cognitive behavior and central and systemic inflammation with aging and sex differences in mice. *Brain, Behavior, and Immunity*, 118:334–354.

Falcon, W. and The PyTorch Lightning team (2023). PyTorch Lightning 2.1.3. Zenodo https://doi.org/10.5281/zenodo.10419201.

Ferri, A., Stagni, V., and Barilà, D. (2020). Targeting the DNA damage response to overcome cancer drug resistance in glioblastoma. *International Journal of Molecular Sciences*, 21(14):4910.

Fey, M. and Lenssen, J. E. (2019). Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Foretz, M., Guigas, B., Bertrand, L., Pollak, M., and Viollet, B. (2014). Metformin: from mechanisms of action to therapies. *Cell metabolism*, 20(6):953–966.

Foretz, M., Guigas, B., and Viollet, B. (2019). Understanding the glucoregulatory mechanisms of metformin in type 2 diabetes mellitus. *Nature Reviews Endocrinology*, 15(10):569–589.

Forster, D. T., Li, S. C., Yashiroda, Y., Yoshimura, M., Li, Z., Isuhuaylas, L. A. V., Itto-Nakama, K., Yamanaka, D., Ohya, Y., Osada, H., et al. (2022). BIONIC: biological network integration using convolutions. *Nature Methods*, 19(10):1250–1261.

Fridley, B. L., Lund, S., Jenkins, G. D., and Wang, L. (2012). A Bayesian integrative genomic model for pathway analysis of complex traits. *Genetic Epidemiology*, 36(4):352–359.

Friedlová, N., Zavadil Kokáš, F., Hupp, T. R., Vojtěšek, B., and Nekulová, M. (2022). Ifitm protein regulation and functions: Far beyond the fight against viruses. *Frontiers in immunology*, 13:1042368.

Gaynanova, I. and Li, G. (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics*, 75(4):1121–1132.

Ge, S. X., Jung, D., and Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, 36(8):2628–2629.

Gentili, M., Martini, L., Sponziello, M., and Becchetti, L. (2022). Biological random walks: multi-omics integration for disease gene prioritization. *Bioinformatics*, 38(17):4145–4152.

Ghahramani, Z. and Jordan, M. (1995). Factorial hidden markov models. In Touretzky, D., Mozer, M., and Hasselmo, M., editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 8.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *International Conference on Machine Learning (ICML)*, pages 1263–1272.

Girden, E. R. (1992). *ANOVA: repeated measures*. Sage.

Goel, S., Singh, R., Singh, V., Singh, H., Kumari, P., Chopra, H., Sharma, R., Nepovimova, E., Valis, M., Kuca, K., et al. (2022). Metformin: Activation of $5'$ AMP-activated protein kinase and its emerging potential beyond anti-hyperglycemic action. *Frontiers in genetics*, 13:1022739.

Goh, W. W. B. and Wong, L. (2016). Evaluating feature-selection stability in next-generation proteomics. *Journal of Bioinformatics and Computational Biology*, 14(05):1650029.

Goldman, C. V. and Zilberstein, S. (2003). Optimizing information exchange in cooperative multi-agent systems. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 137–144.

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology*, 38(6):675–678.

Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268.

Gong, B., Zhou, Y., and Purdom, E. (2021). Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biology*, 22:1–21.

González Olmo, B. M., Bettes, M. N., DeMarsh, J. W., Zhao, F., Askwith, C., and Barrientos, R. M. (2023). Short-term high-fat diet consumption impairs synaptic plasticity in the aged hippocampus via il-1 signaling. *npj Science of Food*, 7(1):35.

Goyal, B., Gill, N. S., Gulia, P., Prakash, O., Priyadarshini, I., Sharma, R., Obaid, A. J., and Yadav, K. (2023). Detection of fake accounts on social media using multimodal data with deep learning. *IEEE Transactions on Computational Social Systems*.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al. (2018). Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422.

Haider, S. P., Burtness, B., Yarbrough, W. G., and Payabvash, S. (2020). Applications of radiomics in precision diagnosis, prognostication and treatment planning of head and neck squamous cell carcinomas. *Cancers of the Head & Neck*, 5(1):1–19.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Hamilton, W. L. (2020). Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159.

Hamilton, W. L., Ying, R., and Leskovec, J. (2018). Representation learning on graphs: methods and applications. *arXiv preprint arXiv:1709.05584*.

Hanahan, D. (2026). Hallmarks of cancer—then and now, and beyond. *Cell*.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *cell*, 100(1):57–70.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, 144(5):646–674.

Hardison, R. C. (2003). Comparative genomics. *PLOS Biology*, 1(2):e58.

Harris, C. R. et al. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hassanzadeh, H. R., Phan, J. H., and Wang, M. D. (2015). A semi-supervised method for predicting cancer survival using incomplete clinical data. In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 210–213.

Hayat, N., Geras, K. J., and Shamout, F. E. (2022). MedFuse: multi-modal fusion with clinical time-series data and chest x-ray images. In *Machine Learning for Healthcare Conference*, pages 479–503. PMLR.

He, H., hao, W., Xi, Y., Chen, Y., Malin, B., and Ho, J. (2024). A flexible generative model for heterogeneous tabular EHR with missing modality. In *International Conference on Learning Representations (ICLR)*.

He, L., Maiolino, P., Leong, F., Lalitharatne, T. D., de Lusignan, S., Ghajari, M., Iida, F., and Nanayakkara, T. (2022). Robotic simulators for tissue examination training with multimodal sensory feedback. *IEEE Reviews in Biomedical Engineering*, 16:514–529.

Hejmadi, M. (2014). *Introduction to cancer biology*. Bookboon.

Hira, Z. M. and Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, 2015:1–13.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the International Conference on Document Analysis and Recognition*, volume 1, pages 278–282. IEEE.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: the next challenge. *Nature Reviews Genetics*, 11(12):855–866.

Hrdlickova, R., Toloue, M., and Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, 8(1):e1364.

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Frontiers in Genetics*, 8:84.

Huang, Z., Zhan, X., Xiang, S., Johnson, T. S., Helm, B., Yu, C. Y., Zhang, J., Salama, P., Rizkalla, M., Han, Z., et al. (2019). SALMON: survival analysis learning with multi-omics neural networks on breast cancer. *Frontiers in Genetics*, 10:166.

Jung, I., Kim, M., Rhee, S., Lim, S., and Kim, S. (2021). MONTI: a multi-omics non-negative tensor decomposition framework for gene-level integrative analysis. *Frontiers in Genetics*, page 1635.

Kamburov, A. and Herwig, R. (2022). ConsensusPathDB 2022: molecular interactions update as a resource for network biology. *Nucleic Acids Research*, 50(D1):D587–D595.

Kang, M., Ko, E., and Mersha, T. B. (2022). A roadmap for multi-omics data integration using deep learning. *Briefings in Bioinformatics*, 23(1):bbab454.

Karczewski, K. J. and Snyder, M. P. (2018). Integrative omics for health and disease. *Nature Reviews Genetics*, 19(5):299–310.

Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, 107(7):2926–2931.

Keicher, M., Burwinkel, H., Bani-Harouni, D., Paschali, M., Czempiel, T., Burian, E., Makowski, M. R., Braren, R., Navab, N., and Wendler, T. (2023). Multimodal graph attention network for COVID-19 outcome prediction. *Scientific Reports*, 13(1):19539.

Keller, A., Eng, J., Zhang, N., Li, X.-j., and Aebersold, R. (2005). A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Molecular Systems Biology*, 1(1):2005–0017.

Khodadadi-Jamayran, A., Akgol-Oksuz, B., Afanasyeva, Y., Heguy, A., Thompson, M., Ray, K., Giro-Perafita, A., Sánchez, I., Wu, X., Tripathy, D., et al. (2018). Prognostic role of elevated mir-24-3p in breast cancer and its association with the metastatic process. *Oncotarget*, 9(16):12868.

Kim, D., Joung, J.-G., Sohn, K.-A., Shin, H., Park, Y. R., Ritchie, M. D., and Kim, J. H. (2015). Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. *Journal of the American Medical Informatics Association*, 22(1):109–120.

Kim, D., Li, R., Dudek, S. M., and Ritchie, M. D. (2013). ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *BioData Mining*, 6(1):1–14.

Kingma, D. P. and Ba, J. (2015). Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Kipf, T. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Kirpich, A., Ainsworth, E. A., Wedow, J. M., Newman, J. R., Michailidis, G., and McIntyre, L. M. (2018). Variable selection in omics data: A practical evaluation of small sample sizes. *PlOS One*, 13(6):e0197910.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Kong, W., Hui, H. W. H., Peng, H., and Goh, W. W. B. (2022). Dealing with missing values in proteomics data. *Proteomics*, 22(23-24):2200092.

Kontomanolis, E. N., Koutras, A., Syllaios, A., Schizas, D., Mastoraki, A., Garmpis, N., Diakosavvas, M., Angelou, K., Tsatsaris, G., Pagkalos, A., et al. (2020). Role of oncogenes and tumor-suppressor genes in carcinogenesis: a review. *Anticancer Research*, 40(11):6009–6015.

Krassowski, M., Das, V., Sahu, S. K., and Misra, B. B. (2020). State of the field in multi-omics research: from computational needs to data mining and sharing. *Frontiers in Genetics*, 11:610798.

Krishnan, P., Kruger, N., and Ratcliffe, R. (2005). Metabolite fingerprinting and profiling in plants using NMR. *Journal of Experimental Botany*, 56(410):255–265.

Krones, F., Marikkar, U., Parsons, G., Szmul, A., and Mahdi, A. (2025). Review of multimodal machine learning approaches in healthcare. *Information Fusion*, 114:102690.

Kuleshov, V., Chaganty, A., and Liang, P. (2015). Tensor factorization via matrix factorization. In *Artificial Intelligence and Statistics*, pages 507–516. PMLR.

Kullmann, S., Heni, M., Hallschmid, M., Fritsche, A., Preissl, H., and Häring, H.-U. (2016). Brain insulin resistance at the crossroads of metabolic and cognitive disorders in humans. *Physiological reviews*.

Kumar, S., Warrell, J., Li, S., McGillivray, P. D., Meyerson, W., Salichos, L., Harmanci, A., Martinez-Fundichely, A., Chan, C. W., Nielsen, M. M., et al. (2020). Passenger mutations in more than 2,500 cancer genomes: overall molecular functional impact and consequences. *Cell*, 180(5):915–927.

Kunej, T. (2019). Rise of systems glycobiology and personalized glycomedicine: why and how to integrate glycomics with multiomics science? *OMICS: A Journal of Integrative Biology*, 23(12):615–622.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., Granton, P., Zegers, C. M., Gillies, R., Boellard, R., Dekker, A., et al. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, 48(4):441–446.

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635.

Lanktree, M. B., Johansen, C. T., Joy, T. R., and Hegele, R. A. (2010). A translational view of the genetics of lipodystrophy and ectopic fat deposition. *Progress in Molecular Biology and Translational Science*, 94:159–196.

Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(1):191–201.

Lee, E. Y. and Muller, W. J. (2010). Oncogenes and tumor suppressor genes. *Cold Spring Harbor Perspectives in Biology*, 2(10):a003236.

Lee, G., Kang, B., Nho, K., Sohn, K.-A., and Kim, D. (2019). MildInt: deep learning-based multimodal longitudinal data integration framework. *Frontiers in Genetics*, 10:617.

Lee, T.-Y., Tseng, C.-J., Wang, J.-W., Wu, C.-P., Chung, C.-Y., Tseng, T.-T., and Lee, S.-C. (2023). Anti-microRNA-1976 as a novel approach to enhance chemosensitivity in XAF1+ pancreatic and liver cancer. *Biomedicines*, 11(4):1136.

Leng, M., Li, Z., Dai, W., and Shi, B. (2024). The power of satellite imagery in credit scoring: a spatial analysis of rural loans. *Annals of Operations Research*, pages 1–38.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. (2018). Feature selection: a data perspective. *ACM Computing Surveys*, 50(6):94.

Li, J., Xu, Y., Peng, G., Zhu, K., Wu, Z., Shi, L., and Wu, G. (2021a). Identification of the nerve-cancer cross-talk-related prognostic gene model in head and neck squamous cell carcinoma. *Frontiers in Oncology*, 11:788671.

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2014). starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research*, 42(D1):D92–D97.

Li, L., Jamieson, K., Rostamizadeh, A., Gonina, E., Ben-Tzur, J., Hardt, M., Recht, B., and Talwalkar, A. (2020). A system for massively parallel hyperparameter tuning. *Proceedings of Machine Learning and Systems*, 2:230–246.

Li, M. M., Huang, K., and Zitnik, M. (2022a). Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering*, 6(12):1353–1369.

Li, N., Zhou, T., and Fei, E. (2022b). Actions of metformin in the brain: a new perspective of metformin treatments in related neurological disorders. *International Journal of Molecular Sciences*, 23(15):8281.

Li, S., Chen, C., and Han, J. (2025). SimMLM: A simple framework for multi-modal learning with missing modality. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24068–24077.

Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J. (2021b). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019.

Liang, P. P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., Chen, L., Wu, P., Lee, M. A., Zhu, Y., et al. (2021). MultiBench: multiscale benchmarks for multimodal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS): Datasets and Benchmarks*, page 1.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., and Stoica, I. (2018). Tune: a research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

Lin, E. and Lane, H.-Y. (2017). Machine learning and systems genomics approaches for multi-omics data. *Biomarker Research*, 5(1):1–6.

Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4):491–502.

Liu, J., Li, S., Cheng, X., Du, P., Yang, Y., and Jiang, W. G. (2019). HOXB2 is a putative tumour promotor in human bladder cancer. *Anticancer Research*, 39(12):6915–6921.

Liu, K., Fu, Y., Wu, L., Li, X., Aggarwal, C., and Xiong, H. (2021). Automated feature selection: A reinforcement learning perspective. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2272–2284.

Liu, L., Fu, Q., Lu, Y., Wang, Y., Wu, H., and Chen, J. (2023). Corrdqn-fs: A two-stage feature selection method for energy consumption prediction via deep reinforcement learning. *Journal of Building Engineering*, 80:108044.

Liu, Q., Cheng, B., Jin, Y., and Hu, P. (2022). Bayesian tensor factorization-drive breast cancer subtyping by integrating multi-omics data. *Journal of Biomedical Informatics*, 125:103958.

Liu, X., Fan, K., Huang, X., Ge, J., Liu, Y., and Kang, H. (2024). Recent advances in artificial intelligence boosting materials design for electrochemical energy storage. *Chemical Engineering Journal*, 490:151625.

Liu, X., Zhang, J., Zhou, S., van der Plas, T. L., Vijayaraghavan, A., Grishina, A., Zhuang, M., Schofield, D., Tomlinson, C., Wang, Y., et al. (2025). Towards deployment-centric multimodal ai beyond vision and language. *Nature Machine Intelligence*, 7:1612–1624.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523.

Lualdi, M. and Fasano, M. (2019). Statistical analysis of proteomics data: a review on feature selection. *Journal of Proteomics*, 198:18–26.

Luo, X., Ju, W., Qu, M., Gu, Y., Chen, C., Deng, M., Hua, X.-S., and Zhang, M. (2022). Clear: Cluster-enhanced contrast for self-supervised graph representation learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Lv, Z., Wang, T., Cao, X., Sun, M., and Qu, Y. (2023). The role of receptor-type protein tyrosine phosphatases in cancer. *Precision Medical Sciences*, 12:57 – 66.

Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., and Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in Biology and Medicine*, 121:103761.

Ma, M., Ren, J., Zhao, L., Testuggine, D., and Peng, X. (2022a). Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186.

Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., and Peng, X. (2021). SMIL: multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310.

Ma, S. (2024). Dataset for "moving towards genome-wide data integration for patient stratification with integrate any omics". Zenodo.

Ma, S., Ren, J., and Fenyö, D. (2016). Breast cancer prognostics using multi-omics data. *AMIA Summits on Translational Science Proceedings*, 2016:52.

Ma, S., Zeng, A. G., Haibe-Kains, B., Goldenberg, A., Dick, J. E., and Wang, B. (2025). Moving towards genome-wide data integration for patient stratification with integrate any omics. *Nature Machine Intelligence*, 7(1):29–42.

Ma, T., Tian, X., Zhang, B., Li, M., Wang, Y., Yang, C., Wu, J., Wei, X., Qu, Q., Yu, Y., et al. (2022b). Low-dose metformin targets the lysosomal ampk pathway through pen2. *Nature*, 603(7899):159–165.

Mamoshina, P., Volosnikova, M., Ozerov, I. V., Putin, E., Skibina, E., Cortese, F., and Zhavoronkov, A. (2018). Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Frontiers in Genetics*, 9:242.

Mansouri, A., Gattolliat, C.-H., and Asselah, T. (2018). Mitochondrial dysfunction and signaling in chronic liver diseases. *Gastroenterology*, 155(3):629–647.

Mantere, T., Tervasmäki, A., Nurmi, A., Rapakko, K., Kauppila, S., Tang, J., Schleutker, J., Kallioniemi, A., Hartikainen, J. M., Mannermaa, A., et al. (2017). Case-control analysis of truncating mutations in DNA damage response genes connects TEX15 and FANCD2 with hereditary breast cancer susceptibility. *Scientific Rep.*, 7(1):681.

Marei, H. E. (2025). Epigenetic regulators in cancer therapy and progression. *NPJ Precision Oncology*, 9(1):1–18.

Mariette, J. and Villa-Vialaneix, N. (2018). Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, 34(6):1009–1015.

McGranahan, N. and Swanton, C. (2017). Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, 168(4):613–628.

Miao, Z., Humphreys, B. D., McMahon, A. P., and Kim, J. (2021). Multi-omics integration in the age of million single-cell data. *Nature Reviews Nephrology*, 17(11):710–724.

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., and Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. *Genes*, 10(2):87.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Mitra, R., McGough, S. F., Chakraborti, T., Holmes, C., Copping, R., Hagenbuch, N., Biedermann, S., Noonan, J., Lehmann, B., Shenvi, A., et al. (2023). Learning from data with structured missingness. *Nature Machine Intelligence*, 5(1):13–23.

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., and Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250.

Mondal, B., Jin, H., Kallappagoudar, S., Sedkov, Y., Martinez, T., Sentmanat, M. F., Poet, G. J., Li, C., Fan, Y., Pruett-Miller, S. M., et al. (2020). The histone deacetylase complex MiDAC regulates a neurodevelopmental gene expression program to control neurite outgrowth. *Elife*, 9:e57519.

Morrice, N., Vainio, S., Mikkola, K., van Aalten, L., Gallagher, J. R., Ashford, M. L., McNeilly, A. D., McCrimmon, R. J., Grosfeld, A., Serradas, P., et al. (2023). Metformin increases the uptake of glucose into the gut from the circulation in high-fat diet-fed male mice, which is enhanced by a reduction in whole-body slc2a2 expression. *Molecular Metabolism*, 77:101807.

Mroueh, Y., Marcheret, E., and Goel, V. (2015). Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134. IEEE.

Newgard, C. B., An, J., Bain, J. R., Muehlbauer, M. J., Stevens, R. D., Lien, L. F., Haqq, A. M., Shah, S. H., Arlotto, M., Slentz, C. A., et al. (2009). A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. *Cell metabolism*, 9(4):311–326.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. (2011). Multimodal deep learning. In *International Conference on Machine Learning (ICML)*, pages 689–696.

Nishiyama, H., Gill, J. H., Pitt, E., Kennedy, W., and Knowles, M. A. (2001). Negative regulation of G(1)/S transition by the candidate bladder tumour suppressor gene DBCCR1. *Oncogene*, 20(23):2956–2964.

Orsolic, I., Carrier, A., and Esteller, M. (2023). Genetic and epigenetic defects of the rna modification machinery in cancer. *Trends in Genetics*, 39(1):74–88.

Ota, J. (2006). Multi-agent robot systems as distributed autonomous systems. *Advanced engineering informatics*, 20(1):59–70.

Pan, Y., Liu, M., Xia, Y., and Shen, D. (2021). Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6839–6853.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Pedregosa, F. et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238.

Perez-Lopez, R., Ghaffari Laleh, N., Mahmood, F., and Kather, J. N. (2024). A guide to artificial intelligence for cancer researchers. *Nature Reviews Cancer*, 24(6):427–441.

Pernicova, I. and Korbonits, M. (2014). Metformin—mode of action and clinical implications for diabetes and cancer. *Nature Reviews Endocrinology*, 10(3):143–156.

Pes, B., Dessì, N., and Angioni, M. (2017). Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data. *Information Fusion*, 35:132–147.

Pettini, F., Visibelli, A., Cicaloni, V., Iovinelli, D., and Spiga, O. (2021). Multi-omics model applied to cancer genetics. *International journal of molecular sciences*, 22(11):5751.

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal*, 19:3735–3746.

Poirion, O. B., Chaudhary, K., and Garmire, L. X. (2018). Deep learning data integration for better risk stratification models of bladder cancer. *AMIA Summits on Translational Science Proceedings*, 2018:197.

Potharlanka, J. L. and M, N. B. (2024). Feature importance feedback with deep q process in ensemble-based metaheuristic feature selection algorithms. *Scientific Reports*, 14(1):2923.

Prasad, D. K. V. and Sitara, R. (2026). Chapter 3 - role of oncogenes, proto-oncogenes, and tumor suppressor genes in cancer progression. In Viswa Prasad, D. K. and Roy, S., editors, *Frontiers of Cancer Biology*, pages 21–26. Academic Press.

Price, E. J., Vitale, C. M., Miller, G. W., David, A., Barouki, R., Audouze, K., Walker, D. I., Antignac, J.-P., Coumoul, X., Bessonneau, V., et al. (2022). Merging the exposome into an integrated framework for "omics" sciences. *iScience*, 25(3):103976.

Raj-Kumar, P.-K., Liu, J., Hooke, J. A., Kovatich, A. J., Kvecher, L., Shriver, C. D., and Hu, H. (2019). PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. *Scientific Reports*, 9(1):7956.

Ramachandram, D. and Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108.

Ramirez, R., Chiu, Y.-C., Hererra, A., Mostavi, M., Ramirez, J., Chen, Y., Huang, Y., and Jin, Y.-F. (2020). Classification of cancer types using graph convolutional neural networks. *Frontiers in Physics*, 8:203.

Rappaport, D. I., Royle, J. A., and Morton, D. C. (2020). Acoustic space occupancy: Combining ecoacoustics and lidar to model biodiversity variation and detection bias across heterogeneous landscapes. *Ecological Indicators*, 113:106172.

Rappoport, N. and Shamir, R. (2019). NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, 35(18):3348–3356.

Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, 49:107739.

Rena, G., Hardie, D. G., and Pearson, E. R. (2017). The mechanisms of action of metformin. *Diabetologia*, 60(9):1577–1585.

Renfrey, S. and Featherstone, J. (2002). Structural proteomics. *Nature Reviews Drug Discovery*, 1(3):175–176.

Repana, D., Nulsen, J., Dressler, L., Bortolomeazzi, M., Venkata, S. K., Tourna, A., Yakovleva, A., Palmieri, T., and Ciccarelli, F. D. (2019). The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biology*, 20:1–12.

Reza, M. K., Prater-Bennette, A., and Asif, M. S. (2024). Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Richesson, R. L., Hammond, W. E., Nahm, M., Wixted, D., Simon, G. E., Robinson, J. G., Bauck, A. E., Cifelli, D., Smerek, M. M., Dickerson, J., et al. (2013). Electronic health records based phenotyping in next-generation clinical trials: a perspective from the nih health care systems collaboratory. *Journal of the American Medical Informatics Association*, 20(e2):e226–e231.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97.

Roche, B., Guégan, J.-F., and Bousquet, F. (2008). Multi-agent systems in epidemiology: a first step for computational biology in the study of vector-borne disease transmission. *BMC bioinformatics*, 9(1):435.

Roman, I., Santana, R., Mendiburu, A., and Lozano, J. A. (2021). In-depth analysis of SVM kernel learning and its components. *Neural Computing and Applications*, 33(12):6575–6594.

Ross, A. (2009). *Fusion, feature-level*, pages 597–602. Springer US, Boston, MA.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Ruan, X., Liu, Y., Wang, P., Liu, L., Ma, T., Xue, Y., Dong, W., Zhao, Y., E, T., Lin, H., et al. (2023). RBMS3-induced circHECTD1 encoded a novel protein to suppress the vasculogenic mimicry formation in glioblastoma multiforme. *Cell Death & Disease*, 14(11):745.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

Ruddon, R. W. (2007). *Cancer biology*. Oxford University Press.

Ruiz, C., Ren, H., Huang, K., and Leskovec, J. (2023). High dimensional, tabular deep learning with an auxiliary knowledge graph. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 26348–26371.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1):3–15.

Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., and Welling, M. (2018). Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer.

Schmitt, L. O. and Gaspar, J. M. (2023). Obesity-induced brain neuroinflammatory and mitochondrial changes. *Metabolites*, 13(1):86.

Schouten, R. M., Lugtig, P., and Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930.

Schulte-Sasse, R. (2020). *Integration of multi-omics data with graph convolutional networks to identify cancer-associated genes*. Freie Universitaet Berlin (Germany).

Schulte-Sasse, R., Budach, S., Hnisz, D., and Marsico, A. (2021). Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nature Machine Intelligence*, 3(6):513–526.

Schwabe, R. F. and Jobin, C. (2013). The microbiome and cancer. *Nature Reviews Cancer*, 13(11):800–812.

Schölkopf, B., Tsuda, K., and Vert, J.-P. (2004). *Support vector machine applications in computational biology*, pages 71–92. MIT Press.

Seoane, J. A., Day, I. N., Gaunt, T. R., and Campbell, C. (2014). A pathway-based data integration framework for prediction of disease progression. *Bioinformatics*, 30(6):838–845.

Sequeira, J. P., Barros-Silva, D., Ferreira-Torre, P., Salta, S., Braga, I., Carvalho, J., Freitas, R., Henrique, R., and Jerónimo, C. (2023). OncoUroMiR: circulating miRNAs for detection and discrimination of the main urological cancers using a ddPCR-based approach. *International Journal of Molecular Sciences*, 24(18):13890.

Setiono, R. and Liu, H. (1997). Neural-network feature selector. *IEEE Transactions on Neural Networks*, 8(3):654–662.

Shang, C., Palmer, A., Sun, J., Chen, K.-S., Lu, J., and Bi, J. (2017). VIGAN: missing view imputation with generative adversarial networks. In *IEEE International Conference on Big Data*, pages 766–775. IEEE.

Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C., and Ester, M. (2019). MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14):i501–i509.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912.

Shur, J. D., Doran, S. J., Kumar, S., Ap Dafydd, D., Downey, K., O'Connor, J. P., Papanikolaou, N., Messiou, C., Koh, D.-M., and Orton, M. R. (2021). Radiomics in oncology: a practical guide. *Radiographics*, 41(6):1717–1732.

Singal, R. and Ginder, G. D. (1999). DNA methylation. *Blood, The Journal of the American Society of Hematology*, 93(12):4059–4070.

Singh, M., Singh, R., and Ross, A. (2019). A comprehensive overview of biometric fusion. *Information Fusion*, 52:187–205.

Skolnick, J., Fetrow, J. S., and Kolinski, A. (2000). Structural genomics and its importance for gene function analysis. *Nature Biotechnology*, 18(3):283–287.

Smit, S., Hoefsloot, H. C., and Smilde, A. K. (2008). Statistical data processing in clinical proteomics. *Journal of Chromatography B*, 866(1-2):77–88.

Sondka, Z., Dhir, N. B., Carvalho-Silva, D., Jupe, S., Madhumita, McLaren, K., Starkey, M., Ward, S., Wilding, J., Ahmed, M., et al. (2024). COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Research*, 52(D1):D1210–D1217.

Song, M., Greenbaum, J., Luttrell IV, J., Zhou, W., Wu, C., Shen, H., Gong, P., Zhang, C., and Deng, H.-W. (2020). A review of integrative imputation for multi-omics datasets. *Frontiers in Genetics*, 11:570255.

Song, T., Zhang, X., Zhang, L., Dong, J., Cai, W., Gao, J., and Hong, B. (2013). miR-708 promotes the development of bladder carcinoma via direct repression of Caspase-2. *Journal of Cancer Research and Clinical Oncology*, 139:1189–1198.

Sood, A., Capuano, A. W., Wilson, R. S., Barnes, L. L., Kapasi, A., Bennett, D. A., and Arvanitakis, Z. (2024). Metformin, age-related cognitive decline, and brain pathology. *Neurobiology of Aging*, 133:99–106.

Spicker, J. S., Brunak, S., Frederiksen, K. S., and Toft, H. (2008). Integration of clinical chemistry, expression, and metabolite data leads to better toxicological class separation. *Toxicological Sciences*, 102(2):444–454.

Stenson, P. D., Ball, E. V., Howells, K., Phillips, A. D., Mort, M., and Cooper, D. N. (2009). The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalised genomics. *Human Genomics*, 4(2):69–72.

Steyaert, S., Pizurica, M., Nagaraj, D., Khandelwal, P., Hernandez-Boussard, T., Gentles, A. J., and Gevaert, O. (2023). Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature Machine Intelligence*, 5(4):351–362.

Stoughton, R. B. (2005). Applications of DNA microarrays in biology. *Annual Review of Biochemistry*, 74:53–82.

Su, X., Hu, P., Li, D., Zhao, B., Niu, Z., Herget, T., Yu, P. S., and Hu, L. (2025). Interpretable identification of cancer genes across biological networks via transformer-powered graph representation learning. *Nature Biomedical Engineering*, pages 1–19.

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, 14:1177932219899051.

Swanton, C., Bernard, E., Abbosh, C., André, F., Auwerx, J., Balmain, A., Bar-Sagi, D., Bernards, R., Bullman, S., DeGregori, J., et al. (2024). Embracing cancer complexity: hallmarks of systemic disease. *Cell*, 187(7):1589–1616.

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., et al. (2023). The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646.

Tabakhi, S., Moradi, P., and Akhlaghian, F. (2014). An unsupervised feature selection algorithm based on ant colony optimization. *Engineering Applications of Artificial Intelligence*, 32:112–123.

Tabakhi, S., Suvon, M. N. I., Ahadian, P., and Lu, H. (2023). Multimodal learning for multi-omics: a survey. *World Scientific Annual Review of Artificial Intelligence*, 1:2250004.

Tan, Z., Jiang, Y., Liang, L., Wu, J., Cao, L., Zhou, X., Song, Z., Ye, Z., Zhao, Z., Feng, H., et al. (2022). Dysregulation and prometastatic function of glycosyltransferase C1GALT1 modulated by cHP1BP3/miR-1-3p axis in bladder cancer. *Journal of Experimental & Clinical Cancer Research*, 41(1):228.

Tang, Y., Chen, D., Wang, L., Zomaya, A. Y., Chen, J., and Liu, H. (2018). Bayesian tensor factorization for multi-way analysis of multi-dimensional EEG. *Neurocomputing*, 318:162–174.

Tao, M., Song, T., Du, W., Han, S., Zuo, C., Li, Y., Wang, Y., and Yang, Z. (2019). Classifying breast cancer subtypes using multiple kernel learning based on omics data. *Genes*, 10(3):200.

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257–284.

Teschendorff, A. E., Jing, H., Paul, D. S., Virta, J., and Nordhausen, K. (2018). Tensorial blind source separation for improved analysis of multi-omic data. *Genome Biology*, 19(1):1–18.

Thaler, J. P., Yi, C.-X., Schur, E. A., Guyenet, S. J., Hwang, B. H., Dietrich, M. O., Zhao, X., Sarruf, D. A., Izgur, V., Maravilla, K. R., et al. (2012). Obesity is associated with hypothalamic injury in rodents and humans. *The Journal of Clinical Investigation*, 122(1):153–162.

The Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315.

The Cancer Genome Atlas Research Network (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, 372(26):2481–2498.

The pandas development team (2021). pandas-dev/pandas: Pandas 1.3.4. Zenodo https://doi.org/10.5281/zenodo.5574486.

The pandas development team (2023). pandas-dev/pandas: Pandas 2.1.1. Zenodo https://doi.org/10.5281/zenodo.8364959.

Theodoridis, S. and Koutroumbas, K. (2008). *Pattern Recognition*. Elsevier Science.

Timper, K. and Brüning, J. C. (2017). Hypothalamic circuits regulating appetite and energy homeostasis: pathways to obesity. *Disease Models & Mechanisms*, 10(6):679–689.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1(Jun):211–244.

Tong, L., Mitchel, J., Chatlin, K., and Wang, M. D. (2020). Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC Medical Informatics and Decision Making*, 20(1):1–12.

Torres-Martos, Á., Anguita-Ruiz, A., Bustos-Aibar, M., Cámara-Sánchez, S., Alcalá, R., Aguilera, C. M., and Alcalá-Fdez, J. (2022). Human multi-omics data pre-processing for predictive purposes using machine learning: a case study in childhood obesity. In *Bioinformatics and Biomedical Engineering*, pages 359–374. Springer.

Tsai, Y.-H. H., Liang, P. P., Zadeh, A., Morency, L.-P., and Salakhutdinov, R. (2019). Learning factorized multimodal representations. In *International Conference on Learning Representations (ICLR)*.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Vangimalla, R. R. and Sreevalsan-Nair, J. (2021). HCNM: heterogeneous correlation network model for multi-level integrative study of multi-omics data for cancer subtype prediction. In *43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1880–1886.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. In *International Conference on Learning Representations (ICLR)*.

Verma, S., Swain, D., Kushwaha, P. P., Brahmbhatt, S., Gupta, K., Sundi, D., and Gupta, S. (2024). Melanoma antigen family A (MAGE A) as promising biomarkers and therapeutic targets in bladder cancer. *Cancers*, 16(2):246.

Virtanen, P. et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272.

Vitrinel, B., Koh, H. W., Kar, F. M., Maity, S., Rendleman, J., Choi, H., and Vogel, C. (2019). Exploiting interdata relationships in next-generation proteomics analysis. *Molecular & Cellular Proteomics*, 18(8):S5–S14.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337.

Wang, D. and Gu, J. (2016). Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quantitative Biology*, 4(1):58–67.

Wang, L., Ding, Z., Tao, Z., Liu, Y., and Fu, Y. (2019a). Generative multi-view human action recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6212–6221.

Wang, M., Wang, K., Liao, X., Hu, H., Chen, L., Meng, L., Gao, W., and Li, Q. (2021a). Carnitine palmitoyltransferase system: a new target for anti-inflammatory and anticancer therapy? *Frontiers in Pharmacology*, 12:760581.

Wang, Q., Chen, R., Cheng, F., Wei, Q., Ji, Y., Yang, H., Zhong, X., Tao, R., Wen, Z., Sutcliffe, J. S., et al. (2019b). A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nature Neuroscience*, 22(5):691–699.

Wang, Q., Zhan, L., Thompson, P., and Zhou, J. (2020). Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1828–1838.

Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., and Huang, K. (2021b). MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12(1):3445.

Ware, A. P., Kabekkodu, S. P., Chawla, A., Paul, B., and Satyamoorthy, K. (2022). Diagnostic and prognostic potential clustered miRNAs in bladder cancer. *3 Biotech*, 12(8):173.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120.

Wen, Y., Song, X., Yan, B., Yang, X., Wu, L., Leng, D., He, S., and Bo, X. (2021). Multi-dimensional data integration algorithm based on random walk with restart. *BMC Bioinformatics*, 22(1):1–22.

Wernstedt, F. (2005). *Multi-agent systems for distributed control of district heating systems*. PhD thesis, Blekinge Institute of Technology.

Wooldridge, M. (2009). *An introduction to multiagent systems*. John Wiley & Sons.

Wörheide, M. A., Krumsiek, J., Kastenmüller, G., and Arnold, M. (2021). Multi-omics integration in biomedical research: a metabolomics-centric review. *Analytica Chimica Acta*, 1141:144–162.

Wu, C.-C., Asgharzadeh, S., Triche, T. J., and D'Argenio, D. Z. (2010). Prediction of human functional genetic networks from heterogeneous data using RVM-based ensemble learning. *Bioinformatics*, 26(6):807–813.

Wu, W., Wang, S., Zhang, Y., Yin, W., Zhao, Y., and Pang, S. (2024a). MOSGAT: uniting specificity-aware GATs and cross modal-attention to integrate multi-omics data for disease diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 28(9):5624–5637.

Wu, X., Yang, L., Wang, J., Hao, Y., Wang, C., and Lu, Z. (2022). The involvement of long non-coding RNAs in glioma: from early detection to immunotherapy. *Frontiers in Immunology*, 13:897754.

Wu, Z., Dadu, A., Tustison, N., Avants, B., Nalls, M., Sun, J., and Faghri, F. (2024b). Multimodal patient representation learning with missing modalities and labels. In *International Conference on Learning Representations (ICLR)*.

Xi, Y., Shen, W., Ma, L., Zhao, M., Zheng, J., Bu, S., Hino, S., and Nakao, M. (2016). HMGA2 promotes adipogenesis by activating C/EBPβ-mediated expression of PPARγ. *Biochemical and Biophysical Research Communications*, 472(4):617–623.

Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., et al. (2025). A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778.

Xu, H., Sang, S., Bai, P., Li, R., Yang, L., and Lu, H. (2023a). GripNet: graph information propagation on supergraph for heterogeneous graphs. *Pattern Recognition*, 133:108973.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019). How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*.

Xu, P., Zhu, X., and Clifton, D. A. (2023b). Multimodal learning with transformers: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132.

Xu, Y., Wang, J., Xu, Y., Xiao, H., Li, J., and Wang, Z. (2017). Screening critical genes associated with malignant glioma using bioinformatics analysis. *Molecular Medicine Reports*, 16(5):6580–6589.

Yamashita, H., Amponsa, V. O., Warrick, J. I., Zheng, Z., Clark, P. E., Raman, J. D., Wu, X.-R., Mendelsohn, C., and DeGraff, D. J. (2017). On a FOX hunt: functions of FOX transcriptional regulators in bladder cancer. *Nature Reviews Urology*, 14(2):98–106.

Yan, K. K., Zhao, H., and Pang, H. (2017). A comparison of graph-and kernel-based omics data integration algorithms for classifying complex traits. *BMC Bioinformatics*, 18(1):1–13.

Yang, B., Yang, Y., and Su, X. (2022). Deep structure integrative representation of multi-omics data for cancer subtyping. *Bioinformatics*, 38(13):3337–3342.

Yang, B., Yang, Y., Wang, M., and Su, X. (2023). MRGCN: cancer subtyping with multi-reconstruction graph convolutional network using full and partial multi-omics dataset. *Bioinformatics*, 39(6):btad353.

Yang, C.-H., Moi, S.-H., Chuang, L.-Y., and Lin, Y.-D. (2025). An information fusion system-driven deep neural networks with application to cancer mortality risk estimate. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):2905–2916.

Yang, Z. and Michailidis, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, 32(1):1–8.

Yao, W., Yin, K., Cheung, W. K., Liu, J., and Qin, J. (2024). DrFuse: learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16416–16424.

Yin, C., Cao, Y., Sun, P., Zhang, H., Li, Z., Xu, Y., and Sun, H. (2022). Molecular subtyping of cancer based on robust graph neural network and multi-omics data integration. *Frontiers in Genetics*, 13:1–14.

Yoon, B.-J. (2009). Hidden markov models and their applications in biological sequence analysis. *Current Genomics*, 10(6):402–415.

You, J., Ma, X., Ding, Y., Kochenderfer, M. J., and Leskovec, J. (2020). Handling missing data with graph representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 19075–19087.

Yuan, Y., Savage, R. S., and Markowetz, F. (2011). Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Computational Biology*, 7(10):e1002227.

Yuan, Z., Li, B., Liao, W., Kang, D., Deng, X., Tang, H., Xie, J., Hu, D., and Chen, A. (2024). Comprehensive pan-cancer analysis of YBX family reveals YBX2 as a potential biomarker in liver cancer. *Frontiers in Immunology*, 15:1382520.

Yun, S., Choi, I., Peng, J., Wu, Y., Bao, J., Zhang, Q., Xin, J., Long, Q., and Chen, T. (2024). Flex-moe: modeling arbitrary modality combination via the flexible mixture-of-experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 98782–98805.

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019). DeepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, 35(24):5191–5198.

Zhang, C., Chu, X., Ma, L., Zhu, Y., Wang, Y., Wang, J., and Zhao, J. (2022a). M3Care: learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2418–2428.

Zhang, C., Song, D., Huang, C., Swami, A., and Chawla, N. V. (2019a). Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 793–803.

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., et al. (2011). International cancer genome consortium fata portal—a one-stop shop for cancer genomics data. *Database*, 2011:bar026.

Zhang, J., Wang, L., Mao, S., Liu, M., Zhang, W., Zhang, Z., Guo, Y., Huang, B., Yan, Y., Huang, Y., et al. (2018a). miR-1-3p contributes to cell proliferation and invasion by targeting glutaminase in bladder cancer cells. *Cellular Physiology and Biochemistry*, 51(2):513–527.

Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., Yuan, D., Tao, Y., Guo, Y., Ni, X., and Shi, T. (2018b). Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. *Frontiers in Genetics*, 9:477.

Zhang, M., Cui, Z., Neumann, M., and Chen, Y. (2018c). An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhang, S., Cooper-Knock, J., Weimer, A. K., Shi, M., Moll, T., Marshall, J. N., Harvey, C., Nezhad, H. G., Franklin, J., dos Santos Souza, C., et al. (2022b). Genome-wide identification of the genetic basis of amyotrophic lateral sclerosis. *Neuron*, 110(6):992–1008.

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*, 40(19):9379–9391.

Zhang, X., Xing, Y., Sun, K., and Guo, Y. (2021). OmiEmbed: a unified multi-task deep learning framework for multi-omics data. *Cancers*, 13(12):3047.

Zhang, X., Zhang, J., Sun, K., Yang, X., Dai, C., and Guo, Y. (2019b). Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 765–769.

Zheng, X., Wang, M., Huang, K., and Zhu, E. (2024). Global and cross-modal feature aggregation for multi-omics data classification and application on drug response prediction. *Information Fusion*, 102:102077.

Zitnik, M., Li, M. M., Wells, A., Glass, K., Gysi, D. M., Krishnan, A., Murali, T., Radivojac, P., Roy, S., Baudot, A., et al. (2024). Current and future directions in network biology. *Bioinformatics Advances*, 4(1):vbae099.

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Information Fusion*, 50:71–91.

Zucknick, M., Richardson, S., and Stronach, E. A. (2008). Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Statistical Applications in Genetics and Molecular Biology*, 7(1).

Appendix A

Additional Biological Findings and Analyses

A.1 Top Biomarkers Identified by HeteroGATomics

Table A.1 Top 30 ranked biomarkers identified by HeteroGATomics in BLCA and LGG. N/A indicates missing biomarker name for the ID.

	BLCA			LGG			
Rank	Biomarker ID	Biomarker name	Omics	Rank	Biomarker ID	Biomarker name	Omics
1	MIMAT0000416	hsa-mir-1-3p	miRNA	1	cg15024277	BEX3	DNA
2	MIMAT0004926	hsa-mir-708-5p	miRNA	2	cg20371266	CHRND	DNA
3	MIMAT0004518	hsa-mir-16-2-3p	miRNA	3	cg22373770	N/A	DNA
4	MIMAT0009451	hsa-mir-1976	miRNA	4	cg05165025	MIDEAS,RP5-1021I20.1	DNA
5	MIMAT0000080	hsa-mir-24-3p	miRNA	5	cg16503259	N/A	DNA
6	DBC1	BRINP1	mRNA	6	cg20253855	CUX1	DNA
7	cg12676289	CTD-2201E9.2,SEMA5A	DNA	7	cg17237063	RBMS3-AS3,RBMS3	DNA
8	cg22777724	HOXB2,HOXB-AS1	DNA	8	cg15275625	N/A	DNA
9	cg26681383	CACNA2D3	DNA	9	RNF126P1	RNF126P1	mRNA
10	cg09313705	HOXB2,HOXB-AS1	DNA	10	TTTY14	TTTY14	mRNA
11	cg20152430	HOXB-AS3,HOXB3	DNA	11	C4orf45	SPMIP2	mRNA
12	LGALS2	LGALS2	mRNA	12	SGCZ	SGCZ	mRNA
13	MAGEA10	MAGEA10	mRNA	13	NAA11	NAA11	mRNA
14	MDH1B	MDH1B	mRNA	14	cg00661753	PTPRA	DNA
15	FOXH1	FOXH1	mRNA	15	BET3L	TRAPPC3L	mRNA
16	YBX2	YBX2	mRNA	16	ZDHHC1	ZDHHC1	mRNA
17	PCDHAC2	PCDHAC2	mRNA	17	G6PC	G6PC1	mRNA
18	LOC644172	LOC644172	mRNA	18	GPR52	GPR52	mRNA
19	CTTNBP2	CTTNBP2	mRNA	19	cg12472597	CTC-1337H24.4,CLCF1,RAD9A,AP003419.11	DNA
20	DMRTA2	DMRTA2	mRNA	20	cg11867599	ABHD18,MFSD8	DNA
21	TEX15	TEX15	mRNA	21	CCL3L1	CCL3L1	mRNA
22	SYBU	SYBU	mRNA	22	cg00020474	N/A	DNA
23	BICC1	BICC1	mRNA	23	cg14302471	LINC00689	DNA
24	FOXA3	FOXA3	mRNA	24	ENC1	ENC1	mRNA
25	cg27452922	N/A	DNA	25	PRR5-ARHGAP8	PRR5-ARHGAP8	mRNA
26	cg11241756	H2BP2	DNA	26	cg14985891	CASQ2	DNA
27	cg20197814	HECW2	DNA	27	cg07971493	MAP3K15	DNA
28	cg00334056	LEMD2	DNA	28	cg08316083	ATP8B3	DNA
29	cg22968622	DND1P1	DNA	29	MIMAT0000707	hsa-mir-363-3p	miRNA
30	cg08836615	N/A	DNA	30	KCNC2	KCNC2	mRNA

A.2 Enrichment Analysis in `HeteroGATomics`

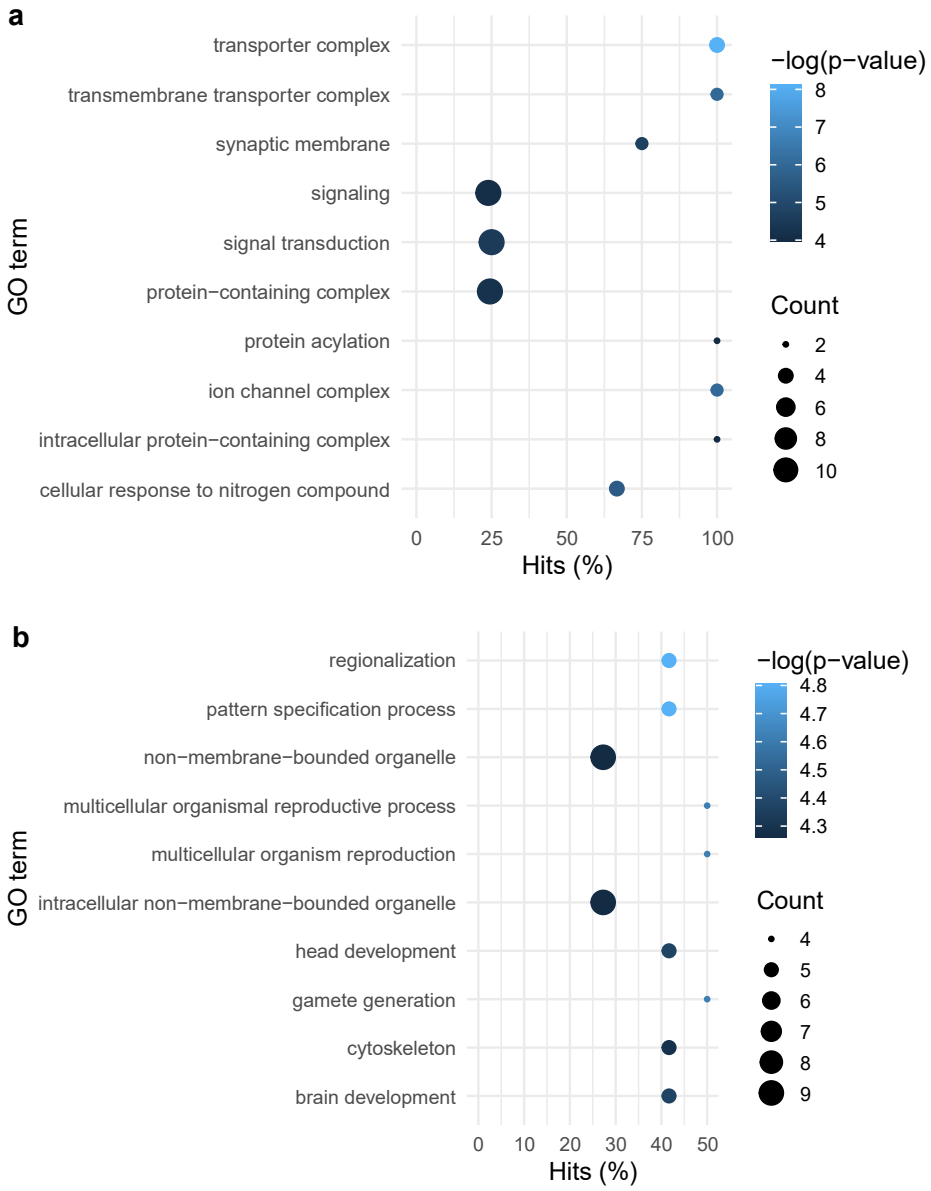

Fig. A.1 **GO enrichment analysis of the top 30 biomarkers from the DNA and mRNA omics modalities.** **a**, Results for the LGG dataset. **b**, Results for the BLCA dataset. The y-axis shows the top 10 most significant GO category terms, while the x-axis represents the percentage of biomarkers belonging to each GO category.

A.3 Interaction Networks in HeteroGATomics

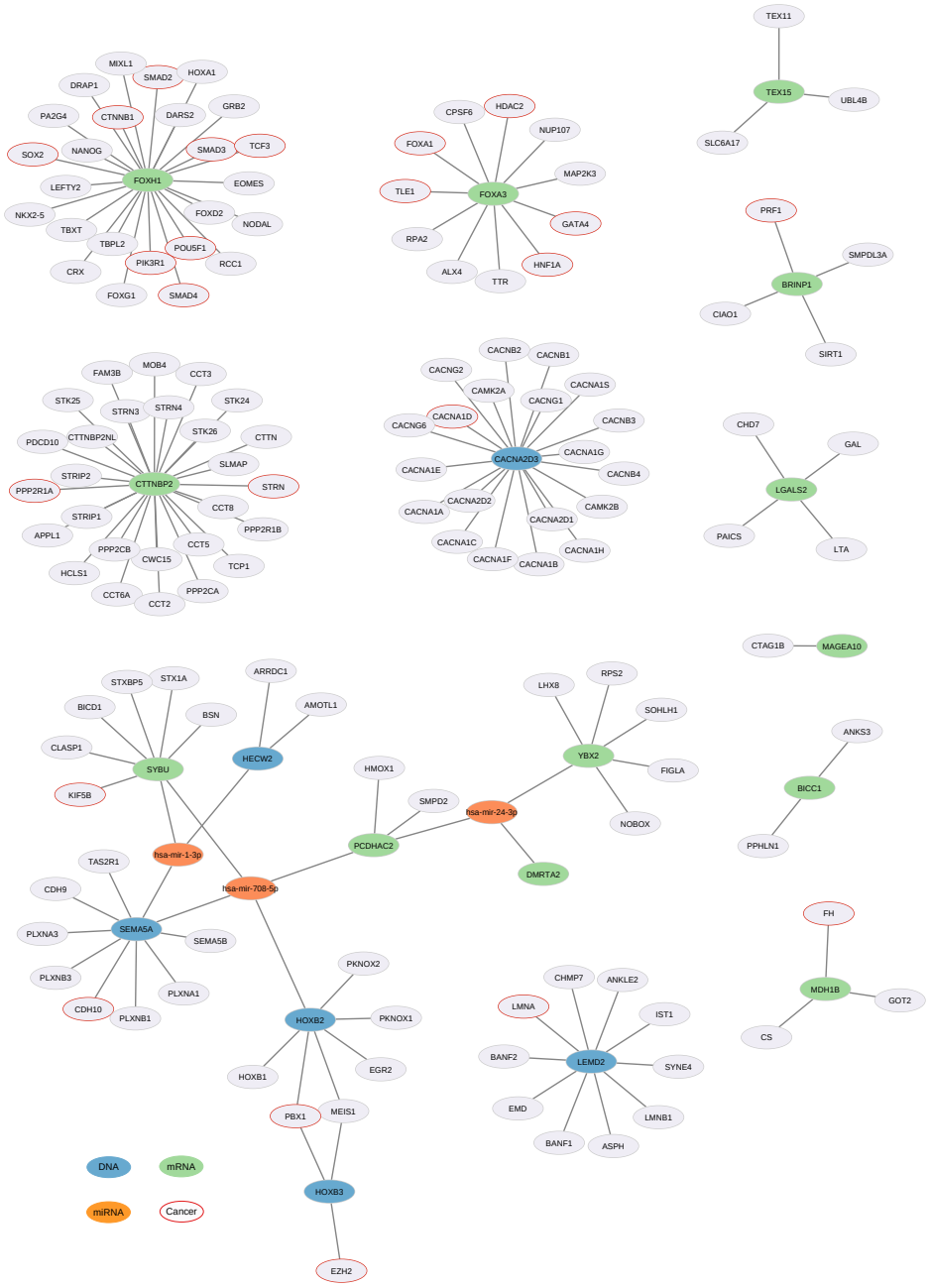

Fig. A.2 **Interaction network of top 30 biomarkers with known partners for BLCA.** Direct protein-protein interactions are recovered for DNA and mRNA omics. For the miRNA omics, known mRNA targets are recovered from starBase (Li et al., 2014). The different omics categories from which the biomarkers originate are indicated as blue (DNA), green (mRNA) and orange (miRNA). Known cancer-related genes are circled in red.

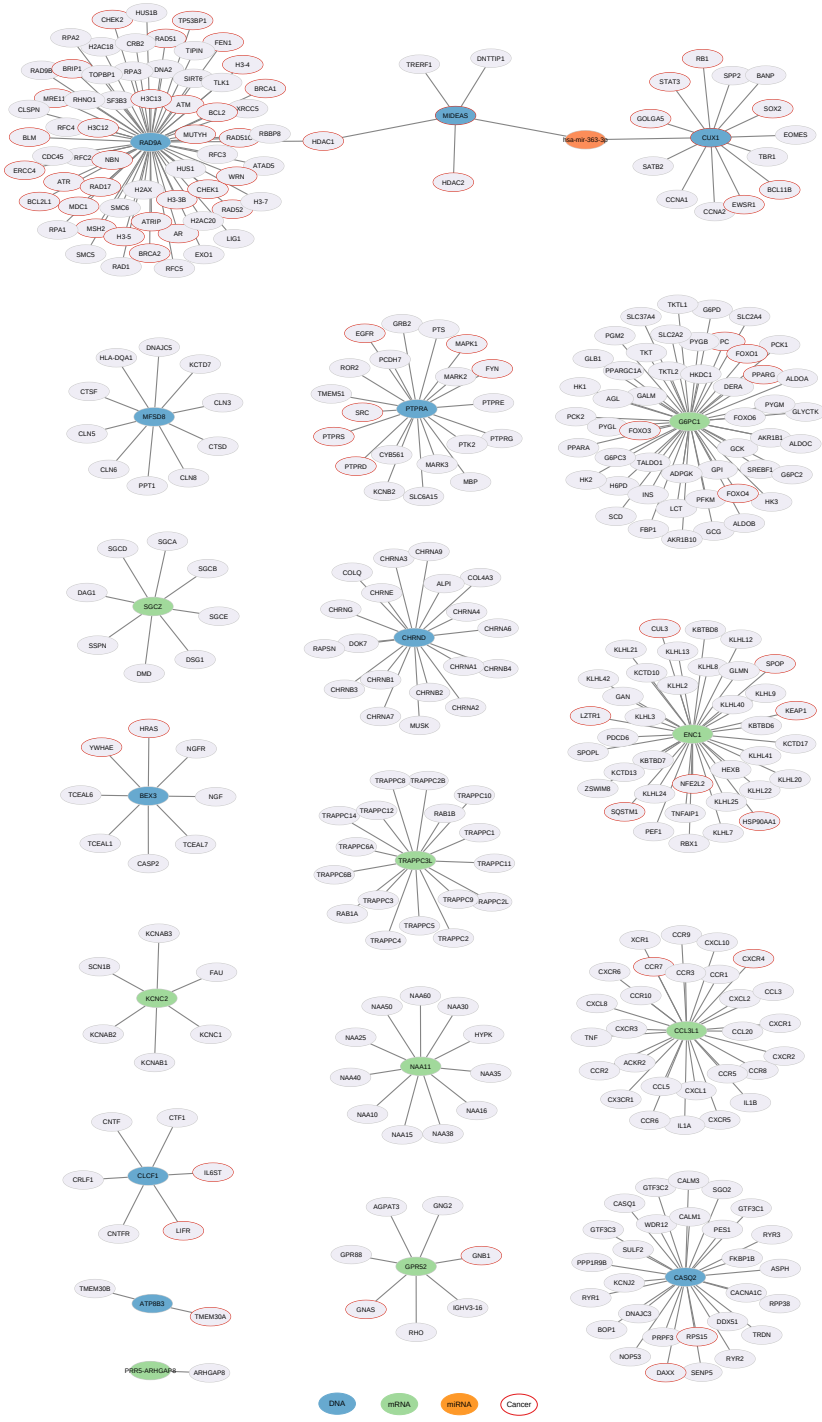

Fig. A.3 **Interaction network of top 30 biomarkers with known partners for LGG.** Direct protein-protein interactions are recovered for DNA and mRNA omics. For the miRNA omics, known mRNA targets are recovered from starBase (Li et al., 2014). The different omics categories from which the biomarkers originate are indicated as blue (DNA), green (mRNA) and orange (miRNA). Known cancer-related genes are circled in red.

A.4 Top-ranked Features Identified by MMAgentOmics

Table A.2 Top coding genes identified by MMAgentOmics using the frequency-based ranking strategy.

Index	Diet factor		Drug factor		Diet and drug factors	
	Score	Gene name	Score	Gene name	Score	Gene name
1	0.4	Ifitm7	1.0	Banp	0.6	Cpt1b
2	0.3	Tcf23	0.8	Efcab6	0.5	Pate2
3	0.3	Trim72	0.7	Zfpm1	0.5	Hmga2
4	0.3	Tmem154	0.6	Snhg12	0.4	Rps12-ps1
5	0.3	Hmga2	0.4	Rax	0.4	Acot3
6	0.2	Cuedc1	0.4	Bhlhe41	0.3	Pah
7	0.2	Adcy4	0.4	Acot3	0.3	Slc2a2
8	0.2	Pate2	0.4	Ext1	0.3	Cdc5lrt9
9	0.2	Msmp	0.4	Sh3bp5	0.3	Efcab6
10	0.2	Mroh2a	0.3	Rcvrn	0.3	Prr5
11	0.1	Ube2c	0.3	Ajap1	0.3	Ms4a7
12	0.1	Crhr2	0.3	Asb11	0.2	AF357399
13	0.1	Hspb7	0.3	Naa10	0.2	Zfhx3
14	0.1	Susd2	0.2	Sgcz	0.2	Rax
15	0.1	Cyba	0.2	Fam186a	0.2	Gprc5d
16	0.1	Mc5r	0.2	Zfhx3	0.2	Prss50
17	0.1	Pax5	0.2	Rnd1	0.2	Zfpm1
18	0.1	Pah	0.2	Igdcc3	0.2	Rpl30-ps8
19	0.1	Pdia2	0.2	Hps4	0.2	Ifi209
20	0.1	Spdef	0.2	Prss50	0.2	Rpl19-ps6
21	0.1	Dpysl4	0.2	Tbc1d1	0.2	Pcdha1
22	0.1	Oxt	0.2	Olfml2b	0.2	Ifitm7
23	0.1	Col11a1	0.1	Itga5	0.2	Banp
24	0.1	Bcar3	0.1	E2f7	0.2	Neurod2

A.5 Enrichment Analysis in MMAgentOmics

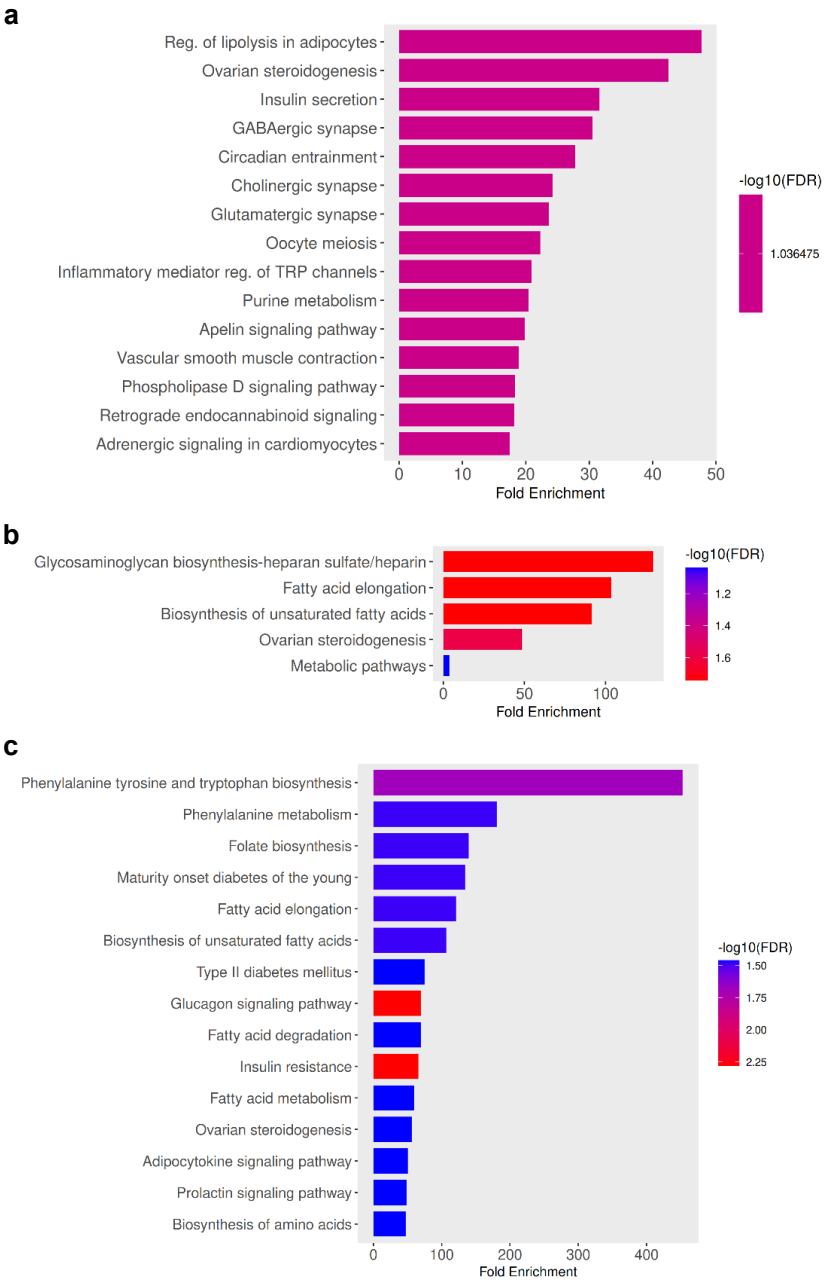

Fig. A.4 **Pathway enrichment analysis of the top coding genes selected from the RNAseq modality by MMAgentOmics. a**, Diet factor (8 genes). **b**, Drug factor (8 genes). **c**, Combined diet and drug factors (7 genes). The y-axis lists the top 15 significantly enriched KEGG pathways, and the x-axis shows their fold enrichment values. Enrichment is performed using ShinyGO v0.85 (Ge et al., 2020) with KEGG as the reference database.

Appendix B

Hyperparameter Settings

B.1 Hyperparameter Tuning for MAGNET

Table B.1 Hyperparameter ranges and selected values across datasets for MAGNET and baseline methods.

Method	Hyperparameter	Range	BRCA	BLCA	OV
MAGNET	MLP Hidden Dimension	Grid Search ([128, 256])	128	256	256
	Patient Graph Sparsity Rate	Discrete Choice ([0.50-0.95], step 0.05)	0.60	0.60	0.80
	Dropout Rate	Discrete Choice ([0.0-0.3], step 0.1)	0.1	0.2	0.2
	#Heads in PMMHA	Grid Search ([2, 4, 8])	2	8	8
	Learning Rate	Log-Uniform (0.00001, 0.001)	0.00032	0.00017	0.000212
MOGONET-Zero	Hidden Dimension	Grid Search ([128, 256])	256	256	128
	#Edges Retained per Node	Integer Uniform ([2-10])	5	4	10
	Dropout Rate	Discrete Choice ([0.0-0.3], step 0.1)	0.3	0.3	0.1
	Pretraining Learning Rate	Log-Uniform (0.00001, 0.001)	0.00076	0.00014	0.000033
	Training Learning Rate	Log-Uniform (0.00001, 0.001)	0.00087	0.001	0.00074
MOGONET-kNN	Hidden Dimension	Grid Search ([128, 256])	128	256	128
	#Edges Retained per Node	Integer Uniform ([2-10])	2	2	2
	Dropout Rate	Discrete Choice ([0.0-0.3], step 0.1)	0.0	0.3	0.2
	Pretraining Learning Rate	Log-Uniform (0.00001, 0.001)	0.00095	0.00029	0.00021
	Training Learning Rate	Log-Uniform (0.00001, 0.001)	0.00083	0.00092	0.0008
MRGCN	#Neighbors per Node	Discrete Choice ([5, 10, 15, 20])	5	15	5
	Dropout Rate	Discrete Choice ([0.0-0.3], step 0.1)	0.2	0.0	0.0
	Learning Rate	Log-Uniform (0.00001, 0.001)	0.00023	0.000024	0.00011
M3Care	Hidden Dimension	Grid Search ([128, 256])	256	256	128
	Dropout Rate	Discrete Choice ([0.0-0.3], step 0.1)	0.2	0.3	0.1
	Learning Rate	Log-Uniform (0.00001, 0.001)	0.0002	0.00041	0.00019
MUSE	Hidden Dimension	Grid Search ([128, 256])	256	256	256
	Dropout Rate	Discrete Choice ([0.0-0.3], step 0.1)	0.3	0.1	0.3
	Learning Rate	Log-Uniform (0.00001, 0.001)	0.00072	0.00018	0.00035