A Phenotype Prediction Framework for Classifying Colorectal Cancer Patients' Response to FOLFOX Treatment: An Integrated Approach



Lebohang Mashatola

Supervisor: Prof. Mandeep Kaur

School of Molecular and Cell Biology
University of the Witwatersrand

This thesis is submitted for the degree of
Doctor of Philosophy

May 2024

"Mohau wa Morena Jesu Kreste, le lerato la Modimo, le kopano ya Moya o Halalelang di be le lona bohle. Amen"

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by the University of the Witwatersrand will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Lebohang Mashatola
May 2024



Acknowledgements

I acknowledge the National Research Foundation and the University of the Witwatersrand for funding this project.

I express my deepest gratitude to my exceptional supervisor, Prof. Mandeep Kaur, whose unwavering support and encouragement have been invaluable throughout my research journey. I am truly fortunate to have had the privilege of working under her mentorship and I sincerely appreciate the patience, belief in me, and inspiration she has provided, shaping not only my academic endeavours but also my personal and professional growth.

I extend my heartfelt gratitude to my friends and colleagues at the school of molecular and cell biology, with special appreciation for Zubayr Kader. In the midst of challenging times, having a friend like you has been invaluable. I would also like to express my gratitude to Dr. Ismail Akhalwaya and Dr. Solomon Assefa for their exceptional mentorship and insightful conversations.

To my fiancée Joy Petersen, thank you for being a constant source of inspiration and encouragement. To my loving parents and sisters, I would not have been able to complete this work without you. Your belief in me has been a constant source of motivation and I dedicate the culmination of this effort to each of you. The journey has been shaped by God's hand, providing me with the courage to face adversities and the wisdom to navigate through life's intricate paths. I am humbled by the continuous blessings and protection that have been bestowed upon me.

Abstract

Colorectal cancer (CRC), characterised by its prevalence and heterogeneity, poses a significant challenge in understanding drug resistance, especially in the context of FOLFOX treatment. This study presents an innovative methodology that integrates diverse data analysis approaches to address the challenge of predicting the phenotype of CRC patients resistant or sensitive to FOLFOX. The initial analysis involved differential and co-expression analyses, identifying pivotal hub genes crucial to drug resistance in CRC, regulating intricate molecular networks. Subsequent enrichment analysis revealed their significant roles in biological functions, particularly influencing DNA repair and nuclear division. To capture inherent topological characteristics within genetic expression data, a novel technique utilising topological data analysis (TDA) was employed. By applying persistence homology to generate persistence images, the Vietoris-Rips complex was constructed using the signed-topological overlap matrix, comprehensively capturing numerous topological features, including high-dimensional *Betti*-1 and *Betti*-2. This provided valuable insights into the structural patterns of gene expression between the hub genes. Furthermore, the integration of whole-slide images enhanced understanding of tissue anatomy, which is crucial for predicting cancer stages. Using a MobileNet architecture, a deep learning model classified cancer stages, contributing to a holistic understanding of colorectal tumor microenvironments. For predictive modelling of drug resistance, a multilayer perceptron applied topological summaries generated by TDA. The developed framework, *GeTopology*, exhibited remarkable performance metrics, achieving an overall 83% accuracy in predicting the FOLFOX response, demonstrating a 3% improvement over a previously published phenotype prediction framework (*NSCLC*) that utilised similar data modes. Robust accuracies were consistently observed in independent datasets, classifying both cancer patients and healthy individuals. The results indicated an approximate 10% increase in model prediction accuracy compared to *NSCLC*, emphasising the potential clinical impact of this integrative approach. In conclusion, this study advances the understanding of drug resistance in CRC by proposing a novel approach that integrates topology with histopathological images, offering transformative insights into predictive modelling and precision medicine.

Scientific Contributions

- The South African Society of Biochemistry and Molecular Biology Conference Conference, Poster Presentation. Title: The Development of a Computational Tool for the Prediction of Cancer Drug Resistance. Authors: Lebohang Mashatola and Mandeep Kaur, Virtual Conference, 23-26 January 2022.
- Published Article: Mashatola, L., Kader, Z., Abdulla, N. et al. Enhancing the Vietoris–Rips simplicial complex for topological data analysis: applications in cancer gene expression datasets. International Journal of Data Science and Analytics (2024). https://doi.org/10.1007/s41060-024-00534-9.

Table of contents

List of figures **x**

List of tables **xii**

Nomenclature **xiii**

Chapter 1 Literature Review **1**

1.1 The Biology of Cancer 1

1.1.1 Hallmarks of Cancer 3

1.1.2 Disease Stages 6

1.1.3 Colorectal Cancer 7

1.1.4 Treatment 8

1.1.5 Drug Resistance 10

1.2 Biological Data 12

1.2.1 High-throughput Sequencing 12

1.2.2 High-throughput Sequencing Data Analysis 17

1.2.3 Medical Image Data 24

1.2.4 Topological Data Analysis 26

1.3 Predictive Modelling 31

1.3.1 Deep Learning 31

1.3.2 Thesis Summary 36

1.4 Aims and Objectives 37

1.4.1 Aims 37

1.4.2 Objectives 37

Chapter 2 Materials and Methods **38**

2.1 Medical Image Data 40

2.1.1 Data Collection 40

2.1.2 Data Processing 41

2.2 Gene Expression Data 44
2.2.1 Data Collection 44
2.2.2 Data Processing 47
2.3 Gene Set Pre-selection 48
2.3.1 Differential Gene Expression Analysis 49
2.3.2 Co-expression Analysis 51
2.3.3 Over-representation Analysis 54
2.4 Data Correlation Analysis 54
2.4.1 Distance Correlations Measures 54
2.4.2 Signed-Topological Overlapping Measures 56
2.5 Persistence Homology 57
2.5.1 Persistence Homology Computation 57
2.5.2 Vector-Based Transformation 58
2.6 Deep Learning 59
2.6.1 Image Classification 59
2.6.2 Topological Summaries Classification 63
2.6.3 Classification using Deep Learning Model using Bimodalities . . 65

Chapter 3 Results **70**
3.1 Differential Gene Expression Analysis 71
3.2 Co-expression Analysis 73
3.3 Topological Summaries 77
3.4 Predictive Learning 81
3.4.1 Image Classification 81
3.4.2 Topological Summaries Classification 82
3.4.3 Classification using Deep Learning Model Built on Bimodalities 83

Chapter 4 Discussion and Conclusions **89**
4.1 Biological Insights into Differentially Expressed Genes and Hub Genes . 91
4.2 Topological Summaries Comparisons 93
4.3 Prediction Learning Comparisons 94
4.4 Limitations of the Present Study 101
4.5 Conclusions 102
4.6 Recommendations for Future Enhancements in the Prediction Models . 104

References **106**

Appendix Chapter A Additional Embodiment's **138**

Appendix Chapter B Running the *GeTopology* Framework 162

List of figures

1.1 The Hallmarks of Cancer . . . 7
1.2 Gene Expression Profiling . . . 16
1.3 Bioinformatics Analysis . . . 17
1.4 RNA-Seq Experimental Workflow and Analysis Pipeline . . . 19
1.5 Sectioned Tumour Sample . . . 27
1.6 *Betti* Numbers . . . 28
1.7 Persistence Homology Computation . . . 30
1.8 Neural Networks . . . 34
1.9 Convolutional Neural Networks . . . 35

2.1 Overall Methodology Performed . . . 39
2.2 Processing Whole-Slide Images . . . 44
2.3 The Process to Identity Hub Genes . . . 48
2.4 Constructing Topological Summaries . . . 60
2.5 Processing and Prediction Learning of WSI . . . 61
2.6 Topological Measures for Predicting FOLFOX Responses . . . 64
2.7 Response to FOLFOX Prediction using WSI and Topological Summaries 66

3.1 Over-representation on Differentially Expressed Genes . . . 73
3.2 Gene Dendrograms . . . 75
3.3 Over-representation of Hub Genes in Gene Modules . . . 76
3.4 Patient Persistence Diagrams . . . 80
3.5 ROC-Curves for WSI Classification . . . 82
3.6 ROC-Curves Comparing Signed-TOM and Distance Correlation Measures 84
3.7 ROC-Curves Comparing *GeTopology* and *NSCLC* . . . 87

4.1 Overview of Findings . . . 90

A.1 PCA Plots . . . 140

A.2 Box-Plots 142
A.3 Density Plots 143
A.4 MA Plots 145
A.5 Enriched KEGG Pathways for DEGs 147
A.6 Soft-thresholding Determination 149
A.7 Intramodular Connectivity 150
A.8 Enriched KEGG Pathways for Hub Genes 152
A.9 Patient Persistence Diagrams for Independent Datasets 155
A.10 ROC-Curves Comparing Signed-TOM and Distance Correlation Measures for Independent Datasets 158
A.11 ROC-Curves Comparing *GeTopology* and *NCSLC* for Independent Datasets 161

B.1 Input Tree Directory for *GeTopology* 164
B.2 Input Tree Directory for Deep Learning on Bimodalities 167

List of tables

1.1 Bioinformatics Analysis Tools 25

2.1 Published Work From The ERI and HCMI-CMDC Projects 41
2.2 Whole-slide Image Metadata 43
2.3 RNA-Seq Count Data Sample Set 45
2.4 Microarray Data Sample Set 46
2.5 *GeTopology* Train and Testing Sample Numbers 67
2.6 Bioinformatics Tools Used 68

3.1 Differentially Expressed Genes 71
3.2 Solt-Thresholding for Gene Expression Data 74
3.3 Hub Gene k_{ME} Values 77
3.4 Over-representation Overlaps 78
3.5 Cancer Stage Prediction Model Performance 81
3.6 Distance Correlation and Signed-TOM for FOLFOX Response Prediction 83
3.7 Performance Metrics for *GeTopology* 85
3.8 *GeTopology* Model Performance Compared to *NSCLC* 86

A.1 Cancer Datasets Obtained from TCGA 153
A.2 Validation Data Phenotype Prediction Evaluation 156
A.3 *GeTopology* and Non-Small-Cell Lung Cancer Classification Tool (*NSCLC*) Comparisons in Independent Cancer Datasets 159

B.1 RNA-Seq Processing Tools using the *GeTopology* Framework 163

Nomenclature

Acronyms / Abbreviations

5-FU 5-Fluorouracil

ABC Adenosine Triphosphate-Binding Cassette

ANN Artificial Neural Network

AUC Area Under the Curve

Bcl-2 B-Cell Lymphoma 2

BCL-XL B-cell Lymphoma-Extra Large

BER Base Excision Repair

BH-Adjusted P-Value Benjamini-Hochberg Probability Value

BiCor Biweight Mid-Correlation

Box and Whisker Plot Box-Plot

BRCA Breast Adenocarcinoma

Caspase Cysteine-Dependent Aspartate-Directed Proteases

CDF Cumulative Distribution Function

cDNA Complementary Deoxyribonucleic Acid

CEL Affymetrix Cell Intensity File

CNN Convolutional Neural Network

COAD Colon Adenocarcinoma

CPD Change Point Detection

CRC Colorectal Cancer

CSC Cancer Stem Cell

CSV Comma Separated Values

CT Computed Tomography

cTNM Clinical Tumour, Nodes and Metastasis Classification

DEG Differentially Expressed Genes

DL Deep Learning

DNA Deoxyribonucleic Acid

E-Cadherin Epithelial-Cadherin

eBayes Empirical Bayesian

EGFR Epidermal Growth Factor Receptor

EMT Epithelial-Mesenchymal Transition

ERI Exceptional Responder Initiative

FC Fold Change

FDA Food and Drug Administration

FDR False Discovery Rate

FN False Negative

FOLFOX Folonic Acid (FOL), Fluorouracil (F) and Oxaliplatin (OX)

FP False Positives

FPKM Fragments Per Kilobase of Feature Per Million

FPM Fragments Per Million

FPR False Positivity Rate

GC-Content Guanine-Cytosine Content

GDC Genomic Data Commons

GO Gene Ontology

GPL Platform

GPU Graphics Processing Unit

GSE Series

GSM Sample

H&E Hematoxylin and Eosin

HCMI-CMDC Human Cancer Model Initiative and the Cancer Model Development Center

KEGG Kyoto Encyclopedia of Genes and Genomes

LUAD Lung Adenocarcinoma

MAPK Mitogen-Activated Protein Kinase

MCL1 Myeloid Cell Leukemia-1

miRNA Micro-Ribonucleic Acid

ML Machine Learning

MLP Multilayer Perceptron

MobileNet Mobile Network

MRI Magnetic Resonance Imaging

mRNA Messenger Ribonucleic Acid

NCBI-GEO Gene Expression Omnibus of the National Centre for Biotechnology Information

NCBI National Center for Biotechnology Information

NCI National Cancer Institute

NER Nucleotide Excision Repair

NHGRI National Human Genome Research Institute

P53 Tumour Protein P53

PCA Principal Component Analysis

PC Principal Component

PCR Polymerase Chain Reaction

PD Persistence Diagram

PH Persistence Homology

PI Persistence Image

PL Persistence Landscape

PNG Portable Network Graphic

Poly(A) Tail Polyadenylation Tail

PRAD Prostate Adenocarcinoma

pTNM Histopathological Tumour, Nodes and Metastasis Classification

PyHIST Histological Image Segmentation Tool

QC Quality Control

RAM Random Access Memory

RB Retinoblastoma

READ Rectum Adenocarcinoma

RECIST Response Evaluation Criteria in Solid Tumours

ReLU Rectified Linear Unit

RMA Robust Multi-array Average

RNA-Seq Ribonucleic Acid-Sequencing

RNA Ribonucleic Acid

ROC-Curve Receiver Operating Characteristic Curve

ROS Reactive Oxygen Species

RPKM Reads Per Kilobase Million

RPM Reads Per Million

Signed-TOM Signed-Topological Overlapping Matrix

SVM Support Vector Machine

SVS Slide Virtual Slide

T-Test T-Statistical Test

TCGA The Cancer Genome Atlas

TDA Topological Data Analysis

TNM Tumour, Nodes and Metastasis

TN True Negative

TP True Positives

TPM Transcripts Per Million

TPR True Positive Rate

TS Thymidylate Synthase

TSV Tab Separated Values

VEGFA Vascular Endothelial Growth Factor A

VR Complex Vietoris–Rips Complex

WGCNA Weighted Gene Co-expression Network Analysis

WSI Whole Slide Image

Chapter 1

Literature Review

1.1 The Biology of Cancer

The central dogma of molecular biology elucidates the transmission of genetic information within a biological system, progressing from deoxyribonucleic acid (DNA) to ribonucleic acid (RNA) and ultimately resulting in the synthesis of functional proteins. Within the cell nucleus, RNA polymerase enzymes transcribe exonic regions of DNA strands to generate RNA, which is then processed to form messenger RNA (mRNA). mRNA constitutes a small percentage (1-3%) of the total RNA present. Subsequently, ribosomes and transfer RNA (tRNA) molecules, acting as adapters, decode mRNA to synthesise specific functional polypeptides. Short mRNA base pair sequences that encode these functional polypeptides are known as genes [1]. It is crucial to emphasise that the regulation of gene expression is fundamental for all biological processes in living organisms.

One such vital process is cell division, which is integral to the growth and maturation of tissues and organs in the body [2]. During this intricate process, DNA replication occurs, which involves duplication of the genome. This replication process can lead to mutations, including single-base pair substitutions and translational frameshifts due to insertions and deletions, which are then transmitted to the resulting daughter cells. Dysregulation of key systems such as mismatch repair, nucleotide excision repair, direct repair, and recombination repair can contribute to the occurrence of DNA mutations. These alterations in DNA sequence have the potential to influence gene expression levels and the activation or deactivation of specific genes [1]. In addition to genetic mutations, environmental factors, such as carcinogens, temperature variations, and external pressures (known as epigenetic effects), can also impact gene expression.

Epigenetic effects involve covalent histone modifications, DNA methylation, and microRNA (miRNA) modulation, all of which contribute to modifying the structure of the chromatin network. The resulting DNA modifications, from mutations and epigenetic alterations, can affect the accessibility of DNA to replication, consequently influencing transcription and translation processes [3]. These alterations have the potential to disrupt normal biological processes, ultimately leading to conditions such as cancer.

Cancer, defined by aberrant cell growth and invasion, is a complex manifestation arising from dysregulation of the meticulously orchestrated processes that govern cell replication and growth. Precise control mechanisms, involving extracellular signalling molecules, tightly regulate cell behaviour, influencing cell size and number. These molecules, which include soluble proteins, cell-binding proteins, and components of the extracellular matrix, play a pivotal role in orchestrating cell behaviour. Mitogens, a subset of these regulatory molecules, act by stimulating cell division, alleviating negative intracellular controls, and influencing the cell cycle. Concurrently, growth factors contribute to cell growth by improving protein synthesis, reducing degradation, and increasing cell mass. Survival factors, on the other hand, prevent apoptosis, a programmed cell death mechanism crucial for maintaining healthy tissue architecture [4]. Apoptosis involves well-defined biochemical events, including cell contraction, nucleic fragmentation, chromatin condensation, DNA fragmentation, and mRNA degradation [4].

Under normal conditions, these regulatory pathways, including tumour suppressor genes, ensure the integrity of tissue architecture and regulate cell numbers, acting as a bulwark against the development of cancer. Tumour suppressor genes, integral to these processes, play a central role in preventing abnormal cell growth and tumorigenesis [5]. The intricate control of cell growth and programmed cell death during normal tissue development and regulation is paramount for understanding the genetic and genomic changes driving cancer [6]. Deregulation of proliferation can lead to conditions such as hyperplasia and neoplasia, while interruptions in cell cycle control points result in genomic damage, reproduction errors, and structural damage. Within this framework, dysregulation of cellular processes gives rise to the hallmarks that characterise the multistep development of human tumours [7]. To understand the complexities of cancer biology, it is imperative to delve into topics such as dysregulation of cell replication and growth, the pivotal roles played by tumour suppressors and oncogenes, and intricate genomic alterations that underlie the initiation and progression of cancer [8]. This

comprehensive overview serves as a foundation for understanding the cumulative effects of genetic and genomic modifications that contribute to the intricate landscape of human tumour development.

1.1.1 Hallmarks of Cancer

Cancer is the process in which healthy cells undergo a transformation into abnormal cells characterised by rapid division and elevated mutation rates, providing them with a survival advantage [9]. The survival edge of cancer cells is facilitated through various mechanisms, collectively known as cancer hallmarks. These hallmarks include proliferative signalling, apoptosis evasion, replicative immortality, escape from growth suppressor signals, angiogenesis, and metastasis [10] (Figure 1.1).

Proliferative Signalling

Maintaining elevated cell proliferation rates is a key characteristic of cancer [11]. In normal cells, carefully regulated production of growth-promoting signals ensures controlled cell division and growth. However, in cancer cells, these signals are dysregulated, often taking the form of abnormal growth factors that bind and activate the tyrosine kinase domains. This activation initiates a cascade of signalling processes that promote cell division and uncontrolled growth [12]. These growth-promoting signals frequently interact with downstream biological pathways involved in cell proliferation, such as the mitogen-activated protein kinase (MAPK) pathway. The MAPK pathway acts as a central hub that integrates various stimuli that influence growth and modulate processes such as cell proliferation, cell death, and cell infiltration. In particular, cancer cells exhibit an increased production of growth factors and tyrosine kinase domains on their cell membranes, increasing their sensitivity to signals driving growth [13]. Furthermore, along with an increased concentration of growth factors, efficient nutrient exchange further supports the accelerated growth of cancer cells.

Sustained Angiogenesis

Ensuring a continuous supply of oxygen and nutrients, along with the elimination of metabolic waste, is crucial to cell survival. Angiogenesis, the process responsible for the formation of the vascular system, plays a key role in meeting these essential needs in both normal tissue and tumour-associated environments [14]. This vascular system consists of endothelial cells that differentiate and organise into blood vessels, which

subsequently sprout to create a network that spans the entire tumour [15]. Central to angiogenesis is the expression of vascular endothelial growth factor A (*VEGFA*), a key factor in this process that is often highly expressed in cancer [16]. Elevated *VEGFA* levels trigger excessive proliferation of endothelial cells, promoting vessel sprouting and resulting in the formation of enlarged blood vessels with increased permeability. Facilitating the efficient exchange of nutrients and waste products not only sustains the increased rate of cell replication but also introduces a layer of complexity that enhances tumour growth. In the midst of this intricate interplay, the quest for replicative immortality takes center stage, as enhanced telomerase expression emerges as a key player, conferring cells with the potential to continuously divide and evade the natural constraints that govern the lifespan of normal cells.

Replicative Immortality

The number of cell divisions in normal cells is inherently restricted to maintain the integrity of normal tissue structure and homeostasis. Telomeric DNA, which consists of repeated sequences, serves a crucial role in preventing the fusion of chromosome ends and in preserving DNA integrity. After each cell division, telomeric DNA undergoes shortening, gradually diminishing chromosomal protection. This reduction in chromosomal protection is associated with a decrease in cell viability and an increase in genomic instability, both integral aspects of the ageing process [17]. In contrast to normal cells, cancer cells employ a distinct strategy to counteract telomeric shortening. In cancer cells, the length of telomeric DNA is maintained through an elevated expression of telomerase, a specialised enzyme that acts as a DNA polymerase by adding repeated sequences to DNA ends. This counteracts the progressive shortening of the telomeres during each round of cell division, contributing to the phenomenon of replicative immortality in cancer cells [18]. Approximately 90% of various types of cancer exhibit increased expression of telomerase, a key factor in the maintenance of replicative immortality [19]. Telomerase, which plays a crucial role in maintaining telomeric DNA length, acts as a guardian of chromosomal protection and stability. While telomerase plays a pivotal role in safeguarding chromosome integrity during DNA replication, the intricate balance of cellular mechanisms extends further to address potential threats through a process known as apoptosis, a programmed cell death pathway crucial for the elimination of damaged cells and the preservation of the overall stability of the organism [17].

Apoptosis Evasion

Programmed cell death through apoptosis serves as a crucial mechanism to control the number of normal and healthy cells, thereby suppressing the development of cancer. Cellular stress, triggered by overstimulation of cancer pathways and the deactivation of cancer suppressing mechanisms, is the primary initiator of apoptosis [20]. Apoptosis orchestration is facilitated by cysteine-dependent aspartate-directed proteases (caspases). Caspases, specifically caspases 8 and 9, are proteolytic enzymes that initiate a series of processes leading to the release of cytochrome c and the breakdown of cellular components [21]. These cascading events result in controlled cell death by cleaving specific proteins in the cytoplasm and nucleus. Counteracting apoptosis are anti-apoptotic proteins, such as B-cell lymphoma (BCL-XL) and myeloid leukemia cell differentiation protein (MCL1). These proteins interfere with the apoptotic process by binding to pro-apoptotic proteins, including Bak and Bax, effectively blocking apoptosis [22]. Cancer cells often increase the expression of anti-apoptotic proteins as a strategy to evade the normal control of cell growth, thereby contributing to the initiation and sustenance of cancer tumours. The hallmark feature of apoptosis evasion in cancer cells, allowing their uncontrolled proliferation and survival, is intricately linked to the dysregulation of key growth inhibitors such as the retinoblastoma protein (RB) and the tumour protein 53 (P53).

Bypassing Growth-Suppressors

Critical growth inhibitors, including RB and P53 respond to external and internal stressors, effectively halting cell growth through mechanisms such as senescence and apoptosis. Cell stressors, such as damaged DNA caused by mutations, insufficient nutrient levels, and low oxygen levels, trigger these inhibitory responses [23]. Despite the presence of these growth-hindering mechanisms, cancer cells often develop strategies to bypass or overcome these controls, allowing them to persist and spread even under unfavourable conditions. This evasion of growth regulation contributes to the progression of the disease, paving the way for more severe and potentially fatal features such as invasion and metastasis [24]. The ability of cancer cells to overcome normal growth limitations underscores the complexity of the disease and the challenges in developing successful treatments that effectively target these aberrant pathways. Understanding and targeting these mechanisms is crucial in the ongoing effort to devise more effective and targeted cancer therapies. Moreover, a deeper exploration into

the molecular intricacies governing metastasis and invasion, critical aspects of cancer progression, becomes imperative for developing approaches that can restrain the spread of malignant cells and improve the overall efficacy of therapeutic interventions.

Invasion and Metastasis

The spread of cancer to distant parts of the body is called metastasis, and this process is closely related to the epithelial-mesenchymal transition (EMT) programme. EMT is responsible for inducing morphological changes in cells, thus altering cell-cell interactions [25]. A key factor in this process is the epithelial cadherin (E-cadherin) cell adhesion molecule, which is crucial for the formation of stable tissue structures in healthy cells. However, in cancer cells, reduced E-cadherin expression increases the likelihood of metastasis [26]. Furthermore, vimentin expression has been identified as a factor that disrupts the mitotic process, resulting in an unstable microtubule system and ultimately leading to a mesenchymal phenotype. This has profound implications for the ability of cells to polarise and migrate effectively [27]. Maintaining cell polarity is essential to regulate cell form, motility, and division, all of which contribute to EMT and increase the propensity to cell invasion [28]. The presence of metastases indicates advanced stages of cancer and is often associated with a poor prognosis. Understanding the molecular mechanisms involved in EMT and metastasis is critical to developing strategies to target and inhibit these processes, potentially improving outcomes for individuals with advanced cancer.

1.1.2 Disease Stages

The likelihood of survival from cancer is affected by various factors, such as the size of the tumour, the duration since the onset of symptoms and the stage of the cancer at the time of diagnosis. The tumour, node, and metastasis (TNM) system, an internationally recognised cancer classification system, is used to categorise cancers based on the size of the primary tumour, the involvement of lymph nodes and the presence of distant metastases [29]. The TNM system comprises two components: the clinical classification (cTNM) and the histopathological classification (pTNM). cTNM staging is used to assess the severity of the cancer before surgery, helping to determine the most suitable treatment. On the contrary, the pTNM staging is used to assess the morphological changes in the cells and nucleus post-surgery, guiding decisions regarding postoperative therapy [30]. The cTNM and pTNM groups are categorised

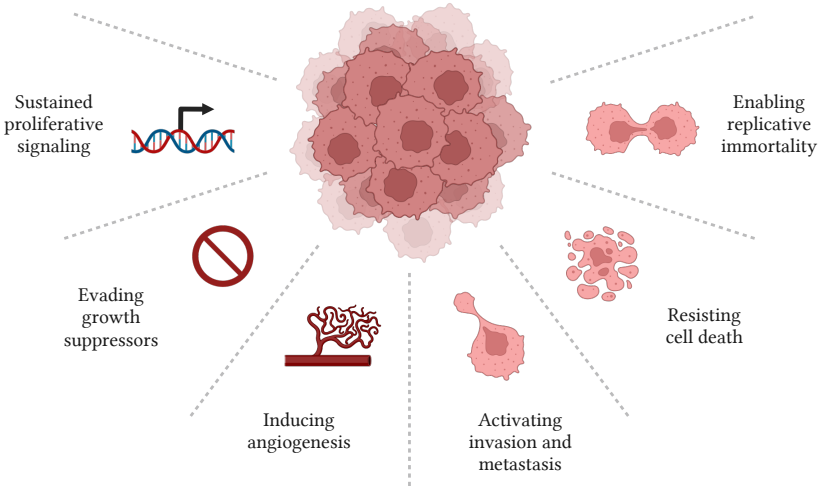


Fig. 1.1 The figure illustrates the hallmarks of cancer, encapsulating key characteristics such as sustained proliferative signalling, evasion of growth suppressors, angiogenesis, invasion and metastasis, evasion of apoptosis, and replicative immortality. These hallmarks collectively define the complex and dynamic nature of cancer progression. The figure was made with biorender.com.

into five prognostic or stage groups. Stage zero cancers lack the ability to spread and exhibit a morphology similar to that of epithelial cells. Stage one cancers are smaller with a low risk of metastasis. Stage two and three cancers share similar characteristics and are associated with larger lymph nodes, indicating the initiation of local invasion through regional node involvement. Stage four cancers are more prone to spreading to other parts of the body and are associated with a poor prognosis [31]. The lethality of cancer, including colorectal cancer (CRC), often depends on the cancer stage, which influences various factors, including tumour characteristics, the presence of metastasis, and the responsiveness of the cancer to available treatments. Early detection and timely intervention can significantly impact outcomes, highlighting the importance of regular screenings and awareness campaigns for CRC and other malignancies.

1.1.3 Colorectal Cancer

CRC holds the global position as the third most common cancer and is the third leading cause of cancer-related deaths in both men and women. According to World

Cancer Research Fund statistics in 2020, CRC ranked third most common cancer worldwide, with approximately 1.9 million new cases reported [32]. However, in South Africa, CRC is ranked as the fourth most prevalent cancer among both sexes [33]. Demographic aging, characterised by an increasing proportion of elderly individuals, significantly contributes to the rising prevalence of CRC. The age-related susceptibility to CRC is linked to genetic and environmental factors, with a notable correlation observed in individuals over 50. Lifestyle factors, such as sedentary habits and poor diet, further compound the risk [34]. Recent findings from [35] highlight a troubling trend: despite being the second most common cause of cancer-related deaths globally, CRC is the leading cause of death in men under 50. This emphasises the urgent need for targeted interventions, including awareness campaigns and early screening initiatives, as demographic shifts continue to shape the landscape of CRC incidence and mortality [36]. Addressing these factors becomes crucial for effective public health strategies and improved outcomes in the face of an aging population.

The five-year survival rate for advanced-stage CRC is significantly lower compared to early-stage CRC, which has an approximate 90% survival rate [37]. However, the incidence and mortality of CRC can be substantially reduced by regular screening methods such as colonoscopy, monitoring, and appropriate treatment [38]. This underscores the critical importance of timely detection and intervention, particularly during the early stages of CRC, to enhance the probability of favourable outcomes.

1.1.4 Treatment

The type of cancer treatment for CRC patients is dependent on the stage and location of the cancer. Treatment options include surgery, chemotherapy, radiation, and immunotherapy. Adjuvant chemotherapy and radiation therapy are often administered after successful surgery or initial treatment to reduce the risk of cancer recurrence [39]. In radiation therapy, high doses are used to effectively destroy cancer cells or impede their growth by causing DNA damage. Cancer cells with irreparable DNA damage cease to divide or die. The body then breaks down and eliminates these targeted cells [40].

In patients diagnosed with early-stage cancer, adjuvant chemotherapy can be beneficial for up to 90% of patients; however, its effectiveness decreases rapidly as the cancer stage progresses. The primary chemotherapy drug used for CRC is 5-fluorouracil

(5-FU) [41]. This drug is generally administered in combination with Oxaliplatin and Leucovorin (Folinic acid), creating a combination therapy known as Folinic acid, Fluorouracil and Oxaliplatin (FOLFOX) [42]. Oxaliplatin induces DNA damage, hindering the production of DNA and RNA. On the other hand, Leucovorin does not have cytotoxic properties; instead, it amplifies the effects of 5-FU by increasing its intracellular concentration [43]. 5-FU is a synthetic analogue of fluorinated pyrimidine, a category of DNA/RNA base pairs that includes cytosine, thymine, and uracil. This synthetic analogue functions by blocking thymidylate synthase (TS), thus preventing the production of the pyrimidine thymidylate essential for DNA replication [44]. For individuals with early-stage CRC treated with 5-FU-based drugs, the 5-year survival rate is approximately 90%, while those with locally advanced CRC have a survival rate of 70%. Unfortunately, individuals with metastatic CRC face a significantly lower 5-year survival rate of 15%, underscoring the critical importance of early detection for CRC, as well as highlighting the influential role of cancer staging in the determination of survival outcomes [45].

Other therapeutic options include monoclonal antibodies that bind and block growth-promoting receptors. For example, Bevacizumab targets VEGFA receptors, which hinder angiogenesis [46]. However, Panitumumab and Cetuximab target the epidermal growth factor receptor (EGFR). In around 80% of CRC cases, both VEGFA receptors and EGFR are overexpressed, making drug treatments effective in reducing angiogenesis and metastasis [47].

Immunotherapy is a recent form of adjuvant therapy that operates by stimulating the immune system within healthy tissue to combat cancer growth and spread. This innovative treatment approach has been associated with a significant improvement in cancer outcomes, marking notable advancements in the field of CRC immunotherapy [48]. In a significant development, the Food and Drug Administration (FDA) has recently approved Pembrolizumab and Nivolumab as second-line therapy for a subset of CRC patients with deficient mismatch repair and microsatellite instability [49]. The combination of chemotherapy and immunotherapy has shown success in increasing patient survival rates. However, a significant challenge in the medical field is that some patients do not benefit from adjuvant therapy, either due to lack of response or the development of resistance over time. Drug resistance poses a crucial issue, leading to poor treatment outcomes and a higher risk of disease recurrence.

1.1.5 Drug Resistance

Cancer cells can develop tolerance to chemotherapy drugs, enabling them to withstand treatments that would typically be lethal. Intrinsic drug resistance is established before any treatment, whereas acquired drug resistance is the result of various factors, including but not limited to the accumulation of genetic and epigenetic alterations in the cancerous tumour [50]. Drug resistance is primarily regulated by both dependent and independent pathways. Dependent pathways activate cell growth, proliferation, and survival, while independent pathways control the formation and maintenance of the tumour microenvironment. These two pathways interact, influencing drug resistance [51]. Consequently, various mechanisms of drug resistance emerge, encompassing modulation of cancer stem cells, drug efflux, genome instability, programmed cell death, and metabolic and immune reprogramming.

Cancer Stem Cells

Cancer stem cells (CSCs) play a significant role in drug resistance, primarily due to their resilience in oxygen-deprived environments. CSCs possess the unique ability to regenerate, differentiate into various cell types, and self-renew [52]. Their inactivity enables them to evade traditional treatments and exhibit enhanced DNA repair, deactivate apoptotic pathways, and effectively manage reactive oxygen species (ROS) and nitrogen species. The molecular processes that underlie the resistance of CSCs to therapy involve maintaining stemness and dormancy, as well as mechanisms such as drug efflux, anti-apoptotic activity, DNA damage repair, and the expansion of the CSC niche [53].

Drug Efflux

Drug resistance is a complex issue, typically involving adenosine triphosphate-binding cassette (ABC) transporter proteins. In particular, the ABCB1, ABCC1, and ABCG2 proteins are strongly associated with resistance to various drugs [54]. Although these transporters are crucial for maintaining healthy tissues by protecting them from xenobiotics and internal metabolites, they are also found in many human tumours, potentially contributing to chemotherapy resistance. Single nucleotide polymorphisms in ABC drug efflux pumps may be linked to drug therapy and disease susceptibility. However, the impact of different genotypes and haplotypes on the expression and activity of these proteins remains unclear, sparking ongoing discussions [55]. Despite efforts

to use efflux pump modulators in cancer treatment, the results have been unsatisfactory, highlighting the complexity of overcoming resistance mechanisms. Genomic instability, a common feature in many cancers, further complicates the efficacy of these modulators as the interplay between genetic alterations and drug resistance mechanisms adds an additional layer of challenge to the development of effective therapeutic strategies.

Genomic Instability

Genomic instability is a hallmark of cancer and has been associated with drug resistance, often arising from DNA mutations, defective DNA repair mechanisms, and chromosomal changes. The treatment with anticancer drugs induces DNA damage, which can inhibit replication, growth, and proliferation. However, DNA damage responses that activate DNA repair mechanisms can enable cancer cells to evade treatment. This is frequently observed in patients treated with Cisplatin and 5-FU [56]. Several DNA repair mechanisms, including nucleotide excision repair (NER), base excision repair (BER), mismatch repair, and non-homologous end joining, can override cell cycle arrest and prevent programmed cell death, ultimately leading to drug resistance [57].

Inhibiting Apoptosis

Chemotherapeutic agents can induce cell death by initiating cytotoxic reactions, such as the generation of ROS, DNA damage, and activation of tumour suppressor genes. This often leads to programmed cell death, primarily through apoptosis, a process triggered by both internal and external pathways. The external pathway involves ligands, cell death receptors (e.g., FAS and tumour necrosis factor receptor-R), linker proteins, and caspases-3, -6, -7, and -8 [58, 59]. The internal pathway, which occurs in mitochondria, involves anti-apoptotic proteins like B-cell lymphoma 2 (Bcl-2) and AKT, as well as pro-apoptotic proteins such as Bax, Bak, and caspase-9. This process results in the breakdown of actin and nuclear lamin proteins, which ultimately leads to cell death [60]. Cancer cells, however, can evade programmed cell death by maintaining high proliferation rates. The increased expression of anti-apoptotic genes and decreased expression of pro-apoptotic genes are strongly associated with drug resistance. Mutations in the TP53 gene, which inhibit apoptosis in response to cell stress and DNA damage, can also contribute to drug resistance by disrupting the link between DNA damage and the initiation of chemotherapeutic agent-triggered apoptosis [61]. In the context of cancer, genetic alterations, abnormal microenvironments,

and treatments can activate pro-survival signaling pathways, rewiring metabolism to enhance survival, growth, and proliferation.

Metabolic Reprogramming

Metabolic reprogramming is a crucial factor in the development of drug resistance in cancer cells. This reprogramming involves more efficient energy utilisation, including processes such as aerobic glycolysis, fatty acid synthesis, and glutaminolysis, while reducing inefficient pathways such as oxidative phosphorylation [62]. Importantly, metabolic reprogramming extends beyond glucose metabolism, also involving cholesterol, lipid, and amino acid metabolism to meet the energy needs of rapidly developing cancer cells. Recent studies have highlighted the role of metabolic reprogramming in cancer cell susceptibility to commonly used first-line chemotherapeutics, emphasising the significant regulatory role of metabolic pathways in resistance to anticancer drugs [63–65]. In the field of drug resistance combating, the integration of biological data and high-throughput sequencing has become indispensable. This approach not only aids in early detection of tumours for preventive measures, but also allows continuous monitoring of cancer patients during treatment. Additionally, the application of new drugs and more effective pharmacological principles is informed by the insights gained from the high-throughput screening of clinical and genomic data, as well as computational modelling. Combining these various strategies for individual patients at different decision-making points can provide a comprehensive understanding and guide the selection of the most effective treatment approaches. One crucial facet of this comprehensive approach involves the analysis of biological data, particularly through high-throughput sequencing.

1.2 Biological Data

1.2.1 High-throughput Sequencing

High-throughput sequencing plays a pivotal role in deciphering the intricate arrangement of base pairs in DNA. This specific order of base pairs constitutes the genetic material known as genes [1]. Initially estimated to be around 20,000, the number of human genes was later revised to approximately 19,890 during the Human Genome Project in 2003 [66]. When DNA sequences are examined, variations in genes can be

discerned between different organisms and species [67].

The advent of high-throughput sequencing has revolutionised our ability to analyse DNA and mRNA sequences on a large scale. This technology allows for the profiling of gene expression data, enabling the simultaneous measurement of the activity of thousands of genes. This comprehensive view of cellular functions facilitates the distinction between actively dividing cells or reveals how cells respond to specific treatments [68]. Many experiments of this nature aim to measure an entire genome simultaneously, assessing each gene present in a particular cell. Various transcriptomics technologies, such as DNA microarrays that measure the pre-identified target genes [69], and sequence-based techniques such as RNA sequencing (RNA-Seq), which provide information on both gene sequences and expression levels, can be used to generate the data required for analysis [70].

DNA Microarrays

A DNA microarray is a collection of microscopic DNA spots attached to a solid surface, typically a microarray chip. Scientists utilise DNA microarrays to measure the expression levels of numerous genes simultaneously or to genotype multiple regions of a genome [71]. The fundamental principle behind microarrays is the hybridisation between two DNA strands—complementary nucleic acid sequences are specifically pair with each other by forming hydrogen bonds between complementary nucleotides (Figure 1.2a).

A higher number of complementary base pairs in a nucleotide sequence results in stronger non-covalent bonding between the two strands. After the non-specific bonding sequences are removed, only stable hybrids remain. Fluorescently labelled target sequences bind to a probe sequence, generating a signal through laser excitation and measurement of the emitted light. The overall signal strength of a spot depends on the quantity of fluorescently labelled molecules present at that spot [69]. Microarrays use relative quantification due to the non-linear nature of fluorescence, making it difficult to accurately determine fluorescence emission yield. Normalisation methods like GeneChip robust multi-array average (gcRMA) and robust multi-array average (RMA) are employed to standardise arrays, allowing for more reliable comparisons between conditions or samples. DNA microarrays can be used to detect RNA, commonly

through complementary DNA (cDNA) after reverse transcription [72].

Due to the biological complexity of gene expression, careful consideration of experimental design is crucial to draw statistically and biologically valid conclusions. Three main elements should be considered when designing a microarray experiment. First, replication of biological samples is essential for robust conclusions. Secondly, technical and biological replicates can aid quantifying precision. Third, fixed probes that bind to target sequences are present as replicates on the microarray slide, providing technical precision in each hybridisation. This precision is feasible only when independently determining the signal of different target sequences. Illumina bead-based arrays, which employ multiple beads with the same probe for a gene, facilitate noise assessment. However, assessing precision becomes more complex when probe sequences differ, as they often vary in yield [73]. Providing information on sample preparation and handling is critical to identify independent units in the experiment and to avoid inflated estimates of statistical significance.

RNA-Seq

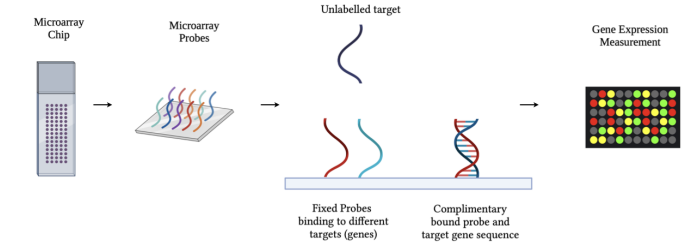
As a successor to microarray technology, RNA-Seq emerged in the mid-1900s to address several limitations inherent in its predecessor. Notably, RNA-Seq overcame challenges associated with cross-hybridisation artefacts, variability in differential expression, and discrepancies in fold change (FC) calculations [70]. The evolution of RNA-Seq can be traced from Sanger sequencing to chemical tag-based techniques. Initially using thin-layer chromatography for tRNA sequencing and primer extension for DNA sequencing; Sanger developed three DNA sequencing methods, while EST sequencing involved sequencing clones from cDNA libraries to the insertion site and SAGE utilised asymmetric restriction endonucleases to generate transcript tags of 12 to 16 nucleotides [74].

The library preparation process begins with the isolation of RNA from target samples, such as cancer tumours. To selectively capture mRNA regions from total RNA, 3' polyadenylation is employed, ensuring that isolated RNA consists predominantly of mature sequences that encode proteins [75]. Isolated mRNA is fragmented into short sequences and reverse transcribed into cDNA. Sequence adapters provide a hybridisation point for sequence primers, allowing extension from adapter sequences. They facilitate identification through additional barcode sequences, and individual

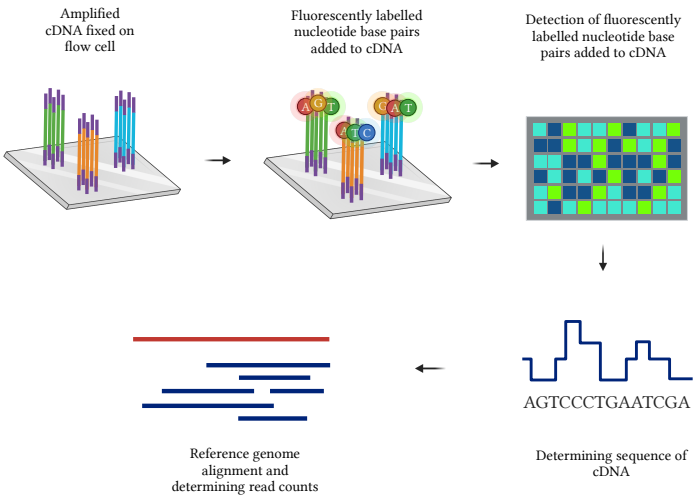
molecule identification is enabled by unique molecular identification (UMI) sequences. These adapters are ligated to cDNA fragments to facilitate sample identification during sequencing. Subsequently, cDNA fragments with adapters undergo polymerase chain reaction (PCR) amplification. During sequencing, the cDNA fragments are immobilised in the RNA-Seq flow cell, and clusters are generated through multiple cycles of bridge amplification. Sequencing by extension involves adding a fluorescently labelled nucleoside triphosphate, which covalently bonds to the extending DNA strand by losing a diphosphate group. The fluorescence signal is detected, followed by chemical cleavage of the fluorophore to remove it, revealing the 3'OH of the deoxyribose for further extension. Subsequently, the fluorophore is washed away, leaving the extending DNA strand available for subsequent reactions. Following each cycle, a washing step is performed before repeating the process [76]. The addition of fluorescently labelled nucleotides can be achieved through single- or paired-end sequencing. Single-end sequencing involves sequencing only one end of a DNA fragment, while paired-end sequencing sequences both ends. The latter allows for the detection of structural changes, such as insertions, deletions, and inversions [77], although it requires more sequencing reads and computational power for analysis [78].

Data analysis in RNA-Seq involves aligning sequenced cDNA fragments with the reference genome. This method uses the count of aligned cDNAs in gene regions to estimate gene expression levels (Figure 1.2b). Short-read RNA-Seq is known for its cost-effectiveness and accuracy in profiling gene expression but has limitations in capturing intricate details of complex gene structures, such as alternative splicing events due to shorter fragment lengths [79]. In contrast, long-read RNA-Seq, such as technologies offered by PacBio and Oxford Nanopore, produces reads that can be the full length of an mRNA species. This approach provides information that is unambiguously from the same molecule, requiring little to no inference. Moreover, unlike short-read approaches where the signal-to-noise ratio attenuates exponentially with length, long-read sequencing, based on single-molecule sequencing, maintains a consistently high signal-to-noise ratio, albeit with higher intrinsic but randomly distributed error. This approach allows direct sequencing of entire transcripts, capturing full-length isoforms and complex gene structures [80]. Despite higher error rates and resource requirements, long-read sequencing excels at revealing detailed information on transcript isoforms, alternative splicing, and structural variations. Effectively managing and analysing the substantial data generated by these high-throughput sequencing techniques is crucial to extracting meaningful insights into the complexities of gene expression and transcriptomic

diversity [81]. Researchers choose between short-read and long-read RNA-Seq based on the study objectives, considering factors such as cost-effectiveness, accuracy, and the level of detail required for their specific transcriptomic analysis [82]. The intricate nature of high-throughput sequencing data requires sophisticated bioinformatics tools and computational approaches for robust analysis and interpretation of the results.



(a) Microarray



(b) RNA-Seq

Fig. 1.2 a) Gene expression profiling using microarray technologies involves utilising a microarray chip with fixed probes for hybridisation with target DNA. The intensity of the signal indicates gene expression levels. Labelling the target DNA on the array makes quantification challenging due to signal production from all targets, even unbound ones. b) In contrast, RNA-Seq involves amplifying cDNA, which binds to a flow cell. Fluorescently labelled nucleotides are added to the cDNA, and their fluorescence indicates quantifiable cDNA fragments. Each sequencing cycle determines the base added to the flow cell spot, enabling straightforward sequence determination without a complex plot. Finally, cDNA fragments are aligned with the reference genome for analysis. The figure was created using biorender.com.

1.2.2 High-throughput Sequencing Data Analysis

Large-scale data analysis without computational approaches is difficult, if not impossible. Figure 1.3 encapsulates the essence of the bioinformatics analysis workflow, outlining the important stages that rely on these computational methods for effective execution and interpretation.

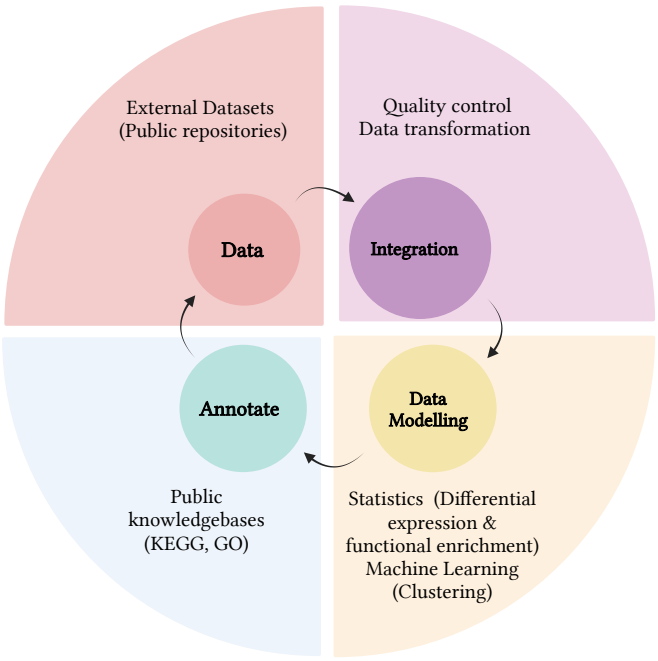


Fig. 1.3 The bioinformatics analysis workflow comprises three main steps: 1) Acquiring data from public repositories including the National Centre for Biotechnology Information (NCBI) and the National Cancer Institute (NCI), 2) Performing quality control (QC) and data transformation, and 3) Utilising high-quality normalised data for statistical data modelling, including differential gene expression analysis (e.g., using GO - Gene Ontology, KEGG - Kyoto Encyclopedia of Genes and Genomes). Subsequently, functional overrepresentation analysis depends on gene annotation conducted using public knowledge bases to establish associations with biological functions. The figure was created using biorender.com.

DNA Microarray Analysis - Normalisation

The analysis of microarray data involves intricate processing of large datasets generated from samples processed on microarray chips. This step is crucial for extracting biological information. While commercial software from manufacturers like Affymetrix

and Agilent is available, open-source tools such as *Limma*, *Oligo*, and *Affy* provide alternative solutions for tasks like background correction and scaling [83–85]. Notably, the robust multi-array average (RMA) method and the removal of signals associated with nonspecific fluorophore bindings enhance result accuracy in this process [86–88].

DNA Microarray Analysis - Identification of Significant Differential Expressions

In the identification of array probes with high or low expressions, various strategies are employed, from simple comparisons to advanced statistical tests such as the T-statistical test (t-test). Interestingly, p-values associated with certain genes may lack reproducibility in replication experiments, emphasising the reliability of lists generated by direct FCs [89]. The FC criteria ensure the identification of genes with significant changes, which contributes to understanding of biological differences between healthy and diseased states [90–92]. This knowledge is essential for interpreting disease states, reflecting the role of differentially expressed genes (DEGs) in RNA-Seq data analysis, which also involves quality control (QC), pre-processing, and downstream differential gene expression analysis (Figure 1.4).

RNA-Seq Analysis - Transcriptome assembly

Pre-processing RNA-Seq data consists of several steps, including reading alignment, transcript quantification, and differential expression analysis. Prior processing ensures that the data is ready for downstream analyses, such as variant calls, co-expression analysis, and functional annotations [93]. The quality of raw sequences is assessed by several steps, including evaluating the measurement of data quality, identifying adapter sequences, and eliminating low-quality measurement. To ensure high-quality data generation, QC is performed at various stages of the RNA-Seq workflow, including sampling, library, and sequence preparation [94]. QC tools such as *Trimmomatic* and *FastQC* are typically used in RNA-Seq workflows [95, 96].

To align short-sequenced cDNA fragments with the transcriptome, alignment to the reference genome is essential. Typically, the alignment algorithm involves two steps: 1) aligning a short part of the reading (the seeds of a genome) and 2) obtaining the optimal alignment through dynamic programming, sometimes combined with known annotations [97]. Various genome alignment software tools are available, including *Bowtie*, *TopHat*, *Subread*, *STAR*, *HISAT2*, and *GMAP-GSNAP* [98–103]. Tools like

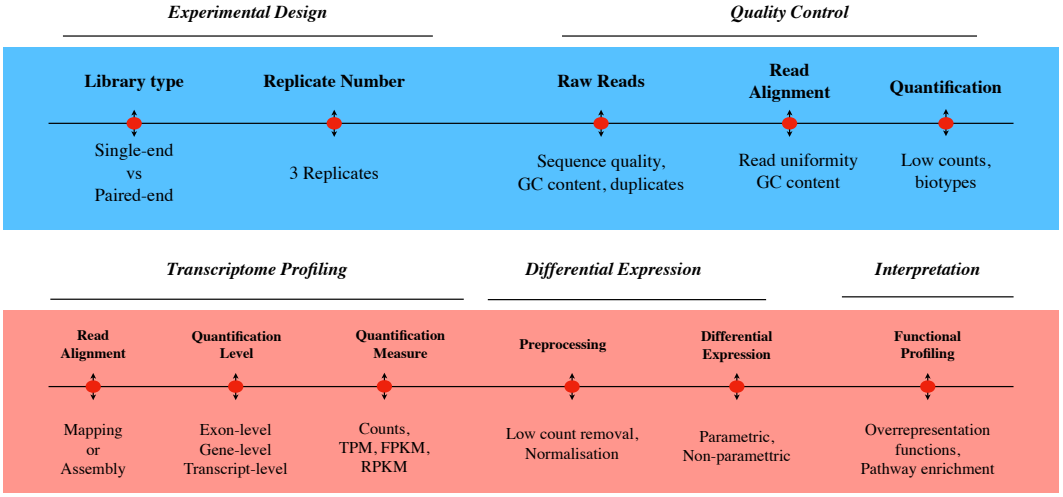


Fig. 1.4 The RNA-Seq experimental workflow and analysis pipeline. The experimental design section outlines the library preparation strategy, incorporating both single-end and paired-end sequencing, with consideration for biological and technical replicates. Quality control steps include the evaluation of raw read sequence quality, read alignment quality, and quantification, focussing on low counts and biotypes. Furthermore, the guanine-cytosine content (GC content) refers to the percentage of nitrogenous bases in a DNA or RNA molecule that consists of either guanine (G) or cytosine (C).The transcriptome profiling phase involves read alignment with the reference genome and quantification at both gene and isoform levels using measures such as Reads Per Kilobase Million (RPKM), Fragments Per Kilobase Million (FPKM), or Transcripts Per Million (TPM). Differential expression analysis encompasses preprocessing steps, including normalisation and filtering, and utilises either parametric or non-parametric methods. The interpretation phase involves overrepresentation functions and pathway overrepresentation analysis to annotate significantly expressed genes for their biological functions and pathway associations. Figure made with biorender.com.

Cufflinks and *StringTie* can be utilised to reconstruct adjacent transcription sequences from the output of genome alignment tools [98, 104, 105].

Moreover, the landscape of cDNA fragment quantification has been dramatically reshaped with the introduction of pseudo-alignment, representing a computationally efficient and highly accurate alternative to conventional methods [106]. Pseudoalignment refers to an efficient algorithm based on k-mers, which utilises a de Bruijn Graph derived from the reference database to identify possible matches for a cDNA sequence

without directly aligning the sample to reference sequences. This approach generates a list of transcripts compatible with each read, avoiding individual base alignment [107]. Pseudoalignment tools, such as *Kallisto* and *Salmon*, minimise the need for large file formats, reduce storage requirements, increase speed, and process large sample amounts with low computing resources [108, 109]. The primary goal of read alignment is to efficiently map short cDNA sequences to large reference genomes and determine the corresponding genomic location from which the short cDNA sequence originates. With the alignment of short cDNA fragments to the reference genome completed, the next critical step in RNA-Seq data analysis involves quantifying gene expression. This process enables the estimation of transcript abundance and provides insights into the dynamic regulation of genes in various biological conditions.

RNA-Seq Analysis - Gene expression quantification

The quantification of cDNA reads mapped during RNA-Seq is crucial for investigating cellular changes induced by external stimuli, disease variations, and various research inquiries. However, accurately estimating gene expression levels faces challenges due to factors such as RNA interference and excessive degradation [110]. To enhance the accuracy of cDNA read estimates, certain parameters need consideration.

Sequencing depth and coverage play pivotal roles in RNA-Seq experiments, and their variations between experiments necessitate normalisation. Converting read counts to fragments per million mapped reads (FPM or RPM) normalises for library size differences, allowing meaningful cross-sample comparisons [111]. The distinction between RPM and FPM arises from the transition from single-end to pair-end sequencing. In single-end sequencing, a fragment equals a read (RPM = FPM), while in pair-end sequencing, two reads represent a fragment (RPM = 2x FPM) [112]. Specifically for RNA-seq, the depth of coverage for a species of RNA depends on factors such as expression level, transcript length, and sequencing depth relative to the complexity of the transcriptome [113].

Consideration of gene length is also essential. Longer transcripts are more likely to be identified than shorter ones, leading to a bias favouring longer genes if expression levels are the same. To address this, FPM is divided by the length of a feature (gene, transcript, or exon), resulting in fragments per kilobase of feature per million mapped reads (FPKM) [114]. For inter-sample comparisons, FPKM is further converted to

transcripts per million (TPM) by dividing each FPKM by the sum of FPKMs within a sample [111]. These normalisation techniques ensure accurate comparisons and meaningful interpretations in various RNA-Seq experiments [113].

The total amount of RNA cannot be accurately determined by evaluating the expression of exons, genes, and transcripts alone. RNA encompasses various species, including structural RNA (such as rRNA, snoRNA, and snRNA), non-coding RNA (such as miRNA and ncRNA), and coding RNA (including primary, partially processed, and completely processed mRNA). Thus, while RNA reflects the expression of genes, assessing total RNA requires considering all RNA species present in the sample. The lower RNA count per gene in samples with a higher total RNA content arises as a result of the dilution effect. When the total amount of RNA is greater, the individual RNA molecules become more dispersed, leading to a reduced concentration of RNA molecules per gene during the sequencing process [115]. These genes appear to have reduced expression, leading to false positive results in downstream analysis [116]. Normalisation strategies, including quantiles and median ratios, attempt to take into account these differences by comparing non-differential expression genes between samples and adjusting their scale accordingly [117]. The total amount of RNA can be established by evaluating the expression of exons, genes, and transcripts. Examining the expression of single exons within a gene is referred to as exon-level quantification and this technique can be used to identify alternative splicing events, which can provide insight into gene regulation and function [118]. Quantifying the expression of a gene as a whole is referred to as gene-level quantification. This method aggregates all transcripts originating from a gene into a single metric, without distinguishing between different alternatively spliced transcripts (isoforms), which may be regulated differently. This approach has the benefit of being more accurate than transcript and exon-level quantification, as it takes into account the entire gene expression profile. Furthermore, it is more efficient, since it requires fewer computational resources [119]. However, gene-level quantification may not be as precise as transcript- and exon-level quantification because it does not consider the individual expression levels of each transcript or exon. Rather than measuring the expression of a gene as a whole, the quantification of transcript levels looks at the expression of individual transcripts within the gene [120]. This approach takes into account the various isoforms of the gene, which may have different expression levels [121]. Furthermore, transcript-level quantification can be more precise in cases where the gene is expressed at low levels or is expressed tissue-specifically [122]. All of these techniques can be used to quantify

total RNA.

In the context of RNA-Seq, variations in gene expression can be modeled using normal, Poisson, or negative binomial distributions, typically categorised as technical and biological variations [123]. RNA-Seq's strength lies in its ability to uncover differences in gene expression between conditions, such as treatment and non-treatment, yielding results commonly referred to as DEGs. These DEGs can be upregulated or downregulated, indicating higher or lower expression levels of interest. Several tools, including *DESeq2*, *EdgeR*, and *Voom+Limma*, are commonly used for differential expression analysis, often implemented using *R/Bioconductor* (available at https://www.bioconductor.org) [124, 125, 83]. Performing differential expression analysis involves considerations such as the input, which contains the RNA-Seq expression matrix (M genes $\times$ N samples) and the design matrix containing the N samples of each experimental condition. The design matrix typically has columns corresponding to the conditional labels that are tested. Additional covariables, including batch effects, known artefacts and metadata, need careful consideration, as they can introduce complexity or mediate gene expression [126]. For human tissue RNA-Seq data, hidden variable analysis is often used to address artifacts not captured in clinical or experimental variables, which are distinct from metadata. Metadata typically describe the data's properties and experimental conditions.

To identify differentially expressed genes (DEGs), researchers have the option to use either parametric or non-parametric techniques. Parametric approaches involve employing statistical models to compare expression levels across different sample sets [127]. For example, the rank-based approach is a parametric method that applies a statistical test to compare gene expression levels between sample groups, ranking the results by significance and identifying the most significantly differentially expressed genes [128]. In contrast, non-parametric approaches do not rely on predefined hypotheses about gene expression patterns [129]. Instead, they fit general smooth spatial patterns of gene expression, requiring fewer stringent assumptions about the data distribution compared to parametric methods. Non-parametric inference has the advantage of being less influenced by outliers, maintaining accuracy and strength even without an outlier filtering step [130]. Many tools leverage regression or non-parametric statistics for identifying DEGs based on various factors such as the number of reference genomes or alignment-free quantification [83, 125, 131–134]. Examples include *DESeq2, Limma, EdgeR, Sleuth, Cuffdiff*, and *Ballgown*. Following regression, these tools commonly

employ false discovery rate (FDR) p-value adjustment to correct for multiple hypotheses [135]. This comprehensive analysis aids researchers in pinpointing genes with significant expression changes, facilitating a deeper understanding of the underlying biological mechanisms.

Typical outputs of a differential expression analysis include rows corresponding to each gene, with at least three columns: the gene's $\log_2$ FC (logFC), p-values, and p-values adjusted for multiple comparisons. Genes are considered biologically significant when they meet the criteria for both effect size (logFC) and statistical significance [136]. These sophisticated methods are designed to mitigate various challenges, such as the increase in false positivite rate (FPR) due to multiple comparisons, sampling preparation artefacts, sample heterogeneity (including mixed genetic backgrounds), highly related samples, incomplete multilevel experiments, and experimental errors [137]. Efforts have been made to compare the results of differential expression analysis tools, and *DESeq2* has shown moderate superiority over other methods [138, 123]. Downstream analysis of DEGs involves two primary approaches: validating observations through the use of technical and biological replicates and drawing biological conclusions by performing overrepresentation analysis. The enrichment of the gene sets determines the statistical significance of the overlap between two sets of genes. This overlap is determined based on pathways or databases, including Gene Ontology (GO) [139] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [140], leading to meaningful biological insights.

Concurrently, co-expression networks act as data-based representations, illustrating genes that exhibit similar behaviour under specific tissue and experimental conditions, further contributing to a comprehensive understanding of the underlying biological processes. They aim to generate hypotheses and association-based approaches to elucidate unknown gene functions [141]. RNA-Seq data are utilised to infer genes in specific pathways based on Pearson, Spearman and distance-based correlation measures, offering the advantage of covering the entire transcriptome for a more comprehensive representation of gene regulation networks. The analysis methodology can detect differential regulation of the same gene sequence isotope, providing information on its biological function [142]. Weighted Gene Co-expression Network Analysis (WGCNA) (Implemented by the R programming package *WGCNA*) is a powerful method for identifying co-expression modules and intramodule hub genes based on RNA-seq data [143, 144]. Hub genes are crucial within complex networks, holding central positions, exerting influence, and coordinating the activities of numerous other genes. Their

centrality underscores their importance in controlling various biological processes [145]. Hub genes often act as integrators of multiple biological pathways, connecting different functional modules and facilitating crosstalk among various cellular processes [146]. Additionally, hub genes contribute to the stability and robustness of biological networks, affecting the structure and function of the network if removed or altered [147]. Hub genes identified through WGCNA can be linked to specific cell types and pathways, with highly connected hubs within each module serving as key regulatory nodes. The eigengene, defined as the weighted sum of all genes expressed in a module, emerges as a valuable biomarker for diagnosis and prediction. A proposed method to estimate the stabilisation coefficient of variation based on RNA-seq data further enhances the robustness of these findings [144]. The bioinformatic tools utilised for data analysis are concisely summarised in Table 1.1.

As mentioned, both microarray and short-read RNA-Seq technologies have inherent imperfections and are prone to bias, which can lead to inaccurate estimates of gene, exon, and transcript abundance [168]. To address these technical limitations and obtain more accurate insights into complex biological phenomena, it is important to implement certain measures. One such strategy is to incorporate technical and biological replicates, which can help validate findings and enhance the reliability of results. However, given the inherent challenges of each technology, a thorough approach involves combining multiple data types. For instance, integrating whole slide images (WSIs) alongside molecular data from microarray and RNA-Seq technologies allows researchers to leverage the strengths of various technologies and compensate for their individual weaknesses. This integrated approach, enriched by the spatial and morphological information provided by WSI, is particularly valuable for understanding intricate biological processes such as drug resistance. In such scenarios, a comprehensive perspective derived from diverse data sources, including WSI, is essential for gaining thorough insights and developing effective solutions.

1.2.3 Medical Image Data

In the contemporary scientific landscape, the advent of digital histopathology stands out as a transformative breakthrough, particularly in the exploration and understanding of various cancers. Histopathology, a cornerstone in the monitoring and surveillance of cancer detection, involves the examination of sectioned biopsies or surgical samples on glass slides [169]. In clinical medicine, pathologists meticulously examine tissue samples,

Table 1.1 Bioinformatics analysis tools categorised by specific tasks for microarray and RNA-Seq data involves addressing critical steps in the analysis pipeline. The tasks encompass data sources, quality control (QC), pre-process, differentially expressed genes (DEG) identification, and gene set analysis (GSA) for both microarray and RNA-Seq data. The comprehensive list of tools with associated references is detailed below. Hyperlinks are provided for each tool or resource mentioned in the table. Tools marked in italics denote C/C++, R, or Python packages, while others indicate web applications or resources. The diverse array of tools selected aims to cover a broad spectrum of functionalities, ensuring a robust and thorough bioinformatics analysis.

Task	**Microarray Data**	**RNA-Seq Data**	**References**
Data Sources	NCBI-GEO, *GEOquery*, Array-express, ENCODE	NCBI-SRA, ENA, DDBJ-DRA	[148–154]
QC	*Limma*	*FastqQC*, *Trimmomatic*, *MultiQC*	[95, 83, 155, 156]
Pre-Process	*Oligo*, *Limma*, *Affy*, *vsn*, *gcrma*	STAR, *HISAT2*, *GMAP-GSNAP*, *Bowtie*, *TopHat*, *Subread*, *SOAPSplice*, *Kallisto*, *Sleuth*, *Salmon*, *HTSeq*, *FeatureCounts*, *Cufflinks*, *Cuffdiff*, *Ballgown*, *StringTie*	[98–104, 108, 109, 122, 83, 132, 84, 85, 157–160]
DEG	*Limma*	*DESeq2*, *EdgeR*	[125, 83, 131]
GSA		DAVID, *ClusterProfiler*, EnrichR, *ReactomePA*, *BiNGO*, GSEA, STRING-db	[161–167]

Note: A hyperlink has been added to each resource in the table. Italicised resources denote C/C++ programs, R, or Python packages, and other resources denote web applications.

choosing the most relevant part for accurate diagnosis of large surgical specimens, which is then meticulously extracted and examined under a microscope after chemical

fixation [170].

For cancer, tissue-based diagnosis is essential for tailoring effective treatment regimens. To visualise the processed specimens under a microscope, sections are stained with various stains to reveal cellular components. The combination of hematoxylin and eosin (H&E) staining is the most commonly used method in histology. Hematoxylin, a basic dye, stains acidic structures such as cell nuclei blue, while eosin, an acidic dye, stains basic structures including the extracellular matrix and connective tissue pink [169] (Figure 1.5). Notably, hematoxylin stains acidic granules in basophilic cells blue due to their acidic nature, while eosin intensity varies based on the basicity of the stained materials, resulting in varying shades of pink. The integration of histopathology with digital pathology, as demonstrated by WSIs, marks a transformative shift in the landscape of cancer diagnosis and treatment planning. WSI, by digitising entire pathology slides, not only enables virtual microscopy but also sets the stage for the application of pioneering technologies such as machine learning (ML) and deep learning (DL) in medical diagnostics [169]. This convergence of traditional histopathological techniques with advanced computational methodologies holds significant promise in improving disease prediction and treatment outcomes. Expanding this paradigm to incorporate gene expression profiling data further enriches the potential for advancements in clinical applications, creating a synergistic approach that takes advantage of both histopathological insights and molecular information for a more comprehensive understanding of cancer biology and improved patient care.

As we explore the intricate landscape of histopathology, a growing field known as topological data analysis (TDA) enters the spotlight. TDA is poised to contribute a unique perspective by capturing the inherent shape and structure within complex biological datasets. By integrating TDA with gene expression profiling data, we can unlock novel insights into the topological features of cancer tissues. This multidimensional approach has the potential to unravel hidden patterns, between gene interactions in the search for more effective cancer diagnostics.

1.2.4 Topological Data Analysis

Examining the shape of the data using geometric and topological principles often aids in the interpretation of complex concepts. In recent years, TDA has emerged as a widely used method for encoding data into shapes, particularly with advances in

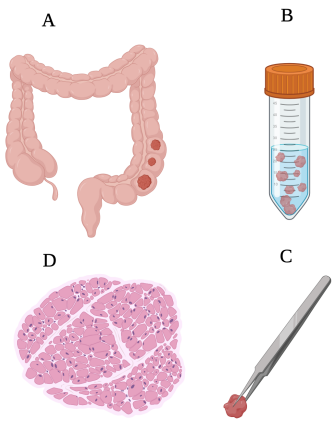


Fig. 1.5 In the process, a tumour sample is first subjected to formaldehyde fixation (A) to preserve its cellular structure. Subsequently, the fixed tumour samples undergo embedding (B) in a suitable medium, followed by microtome sectioning (C) to obtain thin slices for detailed analysis. The obtained sections are then mounted onto slides and stained, typically using H&E staining, a widely utilised histological technique known for revealing cellular components and enhancing the microscopic examination of tissue morphology. The resulting stained tissue sections are then subjected to imaging (D in the current diagram) for further analysis. The figure illustrating these steps was created using biorender.com.

computational geometry facilitated by persistence homology (PH) [171]. In a metric space, objects or data are represented by a collection of points, and the topology is determined by the connections between neighbouring points, either based on correlation or distance [172]. However, topological spaces, which are more general than metric spaces, encode connectivity into a list of open sets. These open sets are subsets of the set of points that describe the topological space [173].

TDA leverages a set of points plotted in a topological space, which is represented by a simplicial complex. In this context, a simplicial complex serves as the foundational structure of the TDA, providing a framework to discern the topological features inherent in the data [174]. The simplicial complex is constructed as a scaffold consisting of simplices—basic elements composed of vertices, points and edges that connect them. These connected vertices collectively form higher-dimensional analogues such as triangles and tetrahedra, capturing intricate connectivity patterns within the data. A key distinction in the construction of simplicial complexes in TDA lies in the choice between the Cech complex and the Vietoris-Rips (VR) complex. The VR complex, a type of simplicial complex, is computationally more expensive than the Cech complex

to compute [175]. It includes all points in the topological space and is constructed by considering pairwise intersections of open sets centred around data points. This computational expense arises due to the exhaustive consideration of all possible pairwise intersections. The choice between Cech and VR complexes involves considerations of computational efficiency, sensitivity to parameter choices, and the desired level of detail in the capture of the underlying topology of the data [176]. Ultimately, the simplicial complex, whether constructed via the Cech or VR complex, becomes a crucial element in applying TDA to gain deeper insights into complex datasets.

Homology groups, represented by *Betti* numbers, form the backbone of TDA, categorising points in the simplicial complex based on their topological features into different homology groups [171]. These groups encapsulate the connectivity and voids within the simplicial complex, providing a robust mathematical foundation for topological analysis [174]. The k^{th}-*Betti* number represents the rank of connectivity of the k^{th} dimensional cycle or the "void" that forms part of the simplicial complex (Figure 1.6).

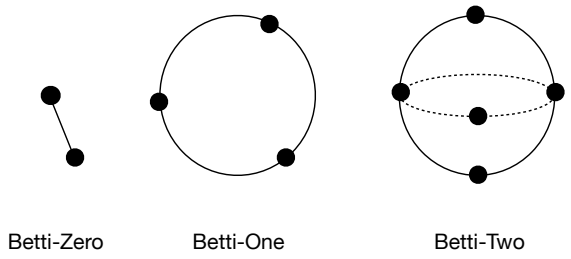


Fig. 1.6 The *Betti* numbers are depicted, each characterising possible shapes within a simplicial complex. A *Betti*-0 topological feature is represented by connected vertices, a *Betti*-1 topological feature is depicted by vertices connected to form an open loop or a cycle, and a *Betti*-2 topological feature is shown as connected vertices forming a three-dimensional sphere or void. The figure was created using biorender.com.

The first three *Betti* numbers include:

- *Betti-0:* Represented by a vertex (or two connected vertices), indicating the count of connected components in the space. In PH, it reveals isolated components that persist on different scales.

- *Betti-1:* Shown as a circle, representing the number of two-dimensional holes or loops in the space. PH tracks the birth and death of these loops on different scales, providing insight into their stability.

- *Betti-2:* Depicted as a sphere, indicating the count of three-dimensional voids or empty regions in the space. In PH, it captures the creation and annihilation of voids at various scales, offering a comprehensive view of the topological structure.

During persistence-based filtration, the timeline of the formation of topological feature sequences is recorded, providing essential information about the *birth* and *death* of each topological feature. In this context, a topological feature is characterised by the filtration value at which it originates (birth coordinate) and the filtration value at which it ceases to exist (death coordinate) [174]. Features that persist through numerous filtration iterations, often termed persistent *Betti* numbers, are considered more informative compared to those appearing for shorter durations, which are more likely to represent noise than meaningful structures [177]. The concept of persistence homology (PH), a mathematical tool within TDA, is based on the persistence of topological features throughout different filtration levels. By systematically recording the birth and death coordinates of these features, PH captures essential aspects of the underlying data structure [171]. Persistent *Betti* numbers, derived from PH, effectively condense intricate data, transforming them into a purely topological representation. These numbers provide valuable information on the longevity and importance of topological characteristics, helping to identify meaningful structures within complex datasets [178]. The *Betti* numbers are typically aggregated into persistence diagrams (PDs), powerful representations encoding the topological features' birth and death coordinates. Each point in a PD encapsulates information about the persistence of a specific topological feature, contributing to a comprehensive understanding of the underlying topological structure of the dataset and these data can be mathematically represented using Equation (1.1).

$$\mathrm{D} = \mathrm{b}_i, \mathrm{d}_i \,|\, \mathrm{b}_i < \mathrm{d}_i \tag{1.1}$$

Where $\mathrm{b_i}$ denotes the birth (the independent variable) and $\mathrm{d_i}$ the death coordinate (the dependent variable) of the topological characteristic k^{th}-*Betti*. PDs embed useful

information for all topological features detected during filtration (Figure 1.7). However, the output is a multiset which cannot be directly implemented in ML and DL machinery. As such, vector-transformation of the PD's multiset is required, and multiple techniques exist for this purpose.

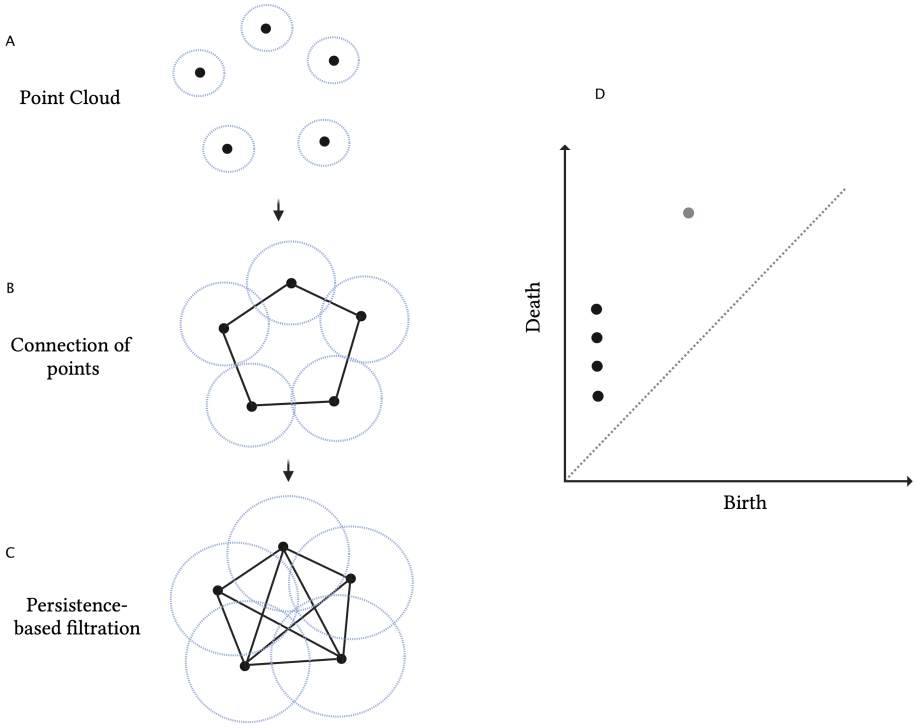


Fig. 1.7 PH computation involves a persistence-based filtration process. In (A), disks centered around each point grow iteratively, and when these disks intersect, an edge is formed. Connected points represent *Betti*-zero topological features, while loops represent *Betti*-one features, as depicted in (B), and (C) illustrates the continuous increase of the disc causing the closing (death) of the loop. In (D), the resulting persistence diagram (PD), with the birth and death axes, illustrates the occurrences of *Betti* numbers with increasing disk sizes. The figure was created using biorender.com..

In the context of TDA, persistent images (PIs) serve as a crucial vector representation of PDs. The most straightforward vector transformation techniques, such as total persistence and persistence entropy, may oversimplify PDs and inadequately characterise data for ML and DL prediction tasks [179]. To address these limitations, more stable and comprehensive vector representations have been used, including persistent landscape (PL) and PI, to improve prediction performance [180, 181]. In this study, we exclude PLs as their consideration is beyond the scope of our investigation. The preference for PIs is grounded in their ability to provide a more concise representation

of PH information, aligning well with the specific research objectives of the present study. The decision to prioritise PIs over PLs is also motivated by their computational efficiency, making them well-suited to the analytical requirements of our research [172].

A Gaussian kernel is used to create PIs from each topological point of the PD, which produces a set of pixels on a fixed rectangular grid. PIs are subsequently formed by pixelating the PD, with the intensity of each pixel representing the weighting function applied to each feature. In particular, more persistent topological characteristics (long lifespan) are considered more significant than less persistent features (short lifespan) [179]. The weight is defined as the absolute value of the lifespan, as indicated by Equation (1.2).

$$\text{Feature} = |\, \mathrm{b}_i - \mathrm{d}_i \,| \tag{1.2}$$

The transformation of high-dimensional tabular data into low-dimensional topological summaries, achieved through PDs, offers a robust approach to encapsulate information about the existence and duration of topological features. These summaries, characterised by birth (b) and death (d) values, enable the representation of complex spatial and structural characteristics in the data. In the realm of ML and DL, the PIs act as valuable input for predictive models, facilitating classification tasks by providing a concise yet informative representation of the underlying data.

1.3 Predictive Modelling

1.3.1 Deep Learning

The information encapsulated in PIs and WSIs can be harnessed through the application of ML and DL models. ML and DL, being mathematical models, are competent in making decisions and predicting outcomes. Standard descriptive statistical models are also mathematical models. What differentiates ML and DL is the learning bases used for attempting to prevent model overfitting, a phenomena that occurs when a ML/DL model learns to perform well on the training data but fails to generalise to unseen data, due to capturing noise or irrelevant patterns in the training data. These algorithms can be broadly categorised into supervised and unsupervised learning

[182]. Supervised learning involves using mathematical functions to identify patterns in labelled data, where each subject in a dataset is assigned an identifier. This identifier can be either a continuous variable representing a range of values that include an infinite number of fractional values, or a discrete variable that signifies a numerical value associated with a specific event. Regression models are commonly used to predict continuous variables, whereas classification models are primarily designed to predict discrete variables. However, it's worth noting that logistic or multinomial regression, despite being categorised as regression models, can be used for classification tasks that predict outcomes on a continuous scale [183]. On the contrary, unsupervised learning is a self-organising approach to learning patterns from unlabelled data, with common applications including data clustering [184]. Supervised models are often considered more interpretable as they can measure and interpret the difference between predicted and actual labels, providing quantifiable results for prediction tasks compared to unsupervised models [185].

Supervised machine learning models face difficulties when dealing with high-dimensional datasets. Nevertheless, the emergence of deep learning, especially artificial neural networks (ANN), has notably enhanced the accuracy of predictive tasks for such data, leading to overall enhancements in predictions. It's crucial to recognise that although deep learning, which includes neural networks, exhibits potential in managing high-dimensional data, the performance of any model hinges on the quality and pertinence of the data it operates on. Inadequate or irrelevant information cannot be compensated for by any mathematical model [186]. These algorithms comprise multiple hidden layers, each housing numerous nodes that compute both linear and non-linear functions. Every node performs a function to calculate a value utilised by the aggregation of layers within the ANN. When nodes are interconnected across all hidden layers, a fully connected neural network is established. Weights are assigned to the edges connecting nodes within a neural network. These weights, which serve as parameters, undergo modification during model training through feedforward, feed-backward or bidirectional processes, evolving over epochs to reflect the iterative nature of learning within the neural network (Figure 1.8) [187].

The feedforward algorithm is based on a set of connected nodes in multiple hidden layers, each connection having distinct weights. These weights assign importance to the input data, with larger weights associated with more significant data and vice versa [188]. During model training, the predicted output is compared to the actual

output to determine the loss or average error. A smaller loss corresponds to better prediction accuracy. Weights are continuously updated in the direction of the derivative, leading to a reduction in prediction loss. This weight update mechanism is known as backpropagation [189]. A multilayer perceptron (MLP), a type of feedforward network, features at least three layers of nodes: the input layer, the hidden layer, and the output layer [190]. All nodes in each layer are connected to every node in the subsequent layer. MLPs are trained using backpropagation, and connections between nodes only flow in one direction, from the input layer to the output layer. While feedforward ANNs are effective in addressing limitations associated with a low number of samples, high data dimensionality, and learning from complex nonlinear data [191], convolutional neural networks (CNNs) offer a specialszed architecture tailored for tasks such as image classification. They utilise convolutional and pooling layers to compress information and capture spatial hierarchies effectively. Convolutional methods essentially reduce dimensions by aggregating features, allowing lower layers to work with more informative summaries, which is critical for complex data processing [192].

CNNs represent a category of neural networks in which hidden layers execute convolutions [193]. The resulting matrices of the input matrix and feature maps are conceptualised as a four-dimensional tensor, comprising the number of input variables, the input height, the input width, and the input channels [194]. CNNs are structured with three key layers: the convolution layer, the pooling layer, and the fully connected layer [195].

In the convolution layer, the input matrix (an image converted into a numerical matrix) is scanned by a kernel or filter, sliding across the image to identify specific features. This process, known as convolution, aims to capture and summarise a region of an image represented by a matrix. Filters, which are two-dimensional arrays of weights that denote particular image regions, may vary in size. The filter is applied to an area of the image and the relationship with the input matrix is computed [192]. A rectified linear unit (ReLU) activation function is then employed to introduce non-linear characteristics to the feature map within the CNN [193].

Following the convolution layer, the pooling layer reduces the dimensionality of the input matrix by iteratively traversing the entire input matrix. In the max-pooling process, the pooling filter selects a pixel within its region and outputs the maximum value of pixels to the output array [196]. Finally, a fully connected layer connects

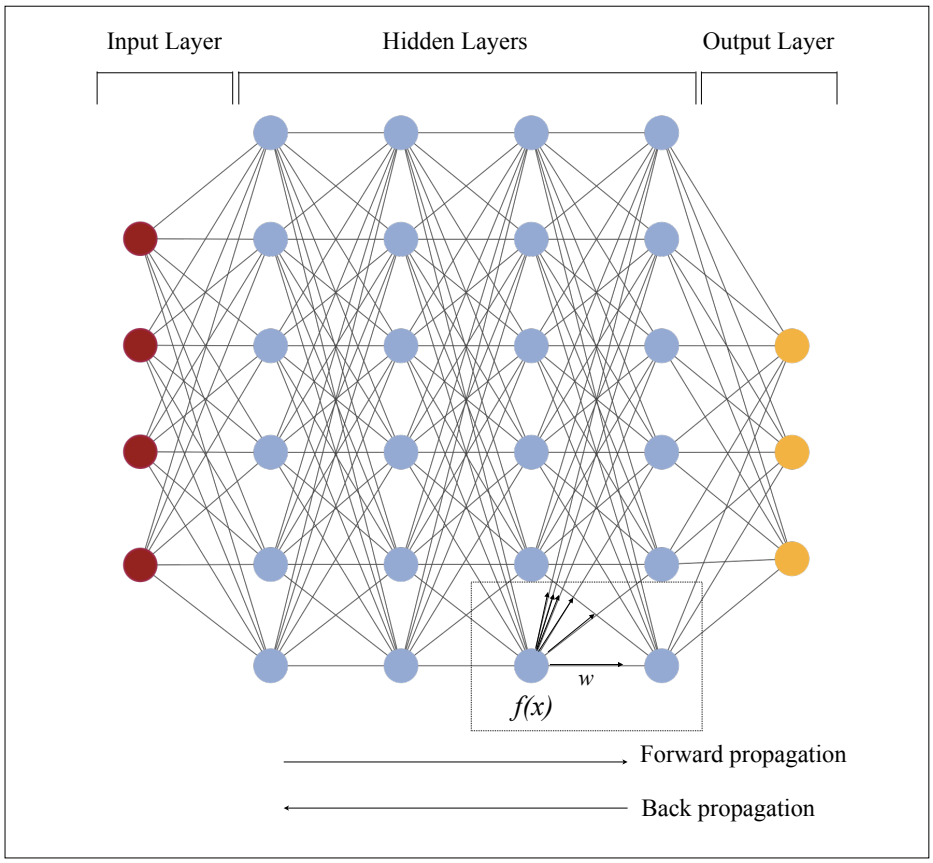


Fig. 1.8 A neural network depicting the feedforward algorithm, where input data traverse through multiple nonlinear parameters represented as nodes within the hidden layer (The arrows connecting the nodes in one of the hidden layers illustrate the flow of data from the input nodes, through the hidden nodes, and to the output nodes in a forward direction). The weights are initialised, and the predicted value is computed. Subsequently, the predicted value is compared with the actual value to measure the loss. To enhance the model's performance, backpropagation is executed, adjusting and updating the weights for each parameter to minimise the loss. The model undergoes re-training by forwarding the data through the network with the updated weights, and the loss is computed. This iterative process of forward- and backpropagation continues until the loss function and prediction accuracy are optimised, resulting in a finely tuned neural network. The colour-coded nodes visually distinguish the input layer in red, hidden layers in blue and the output layer in yellow throughout the learning process. The figure was created using biorender.com.

each node in the output directly to every node in the preceding layer, defining a fully connected CNN (Figure 1.9). CNNs, widely applied in the classification of image data,

find further utility when integrated into a broader framework of multimodal DL.

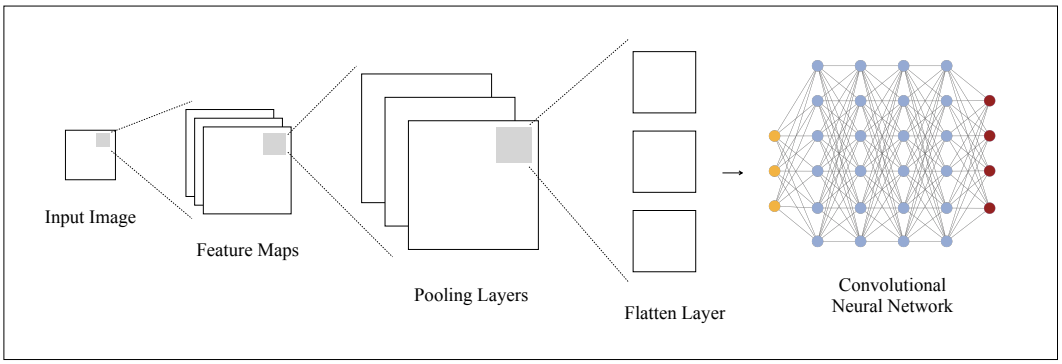


Fig. 1.9 A Convolutional Neural Network (CNN) undergoes a sequence of processes involving the transformation of the image into feature maps, pooling layers, and flattened layers through convolutional operations. These processes aim to systematically capture distinctive regions within the input image. The figure was created using biorender.com.

Leveraging DL from multiple modalities of data provides a powerful strategy to enhance prediction learning by capturing complementary information that may remain undetected by a single data modality [197]. The term "bimodality" refers to the use of two distinct data modalities, such as images and pathological information, to enrich the understanding of a given task. Unlike traditional ML and DL tasks that often concentrate on a singular modality of data for predictions, incorporating multiple signals, empowers prediction models to make more well-informed decisions [198]. This inclusive approach allows for a more comprehensive exploration of the complexities in the data and fosters a deeper understanding of the intricacies of the task at hand [197]. The unexplored synergy between topological summaries in the form of PIs and WSI data represents a significant gap in current understanding of resistance to FOLFOX. Recognising this gap, our study aims to bridge the gap by employing DL learning, harnessing the complementary strengths of both data modalities. By doing so, we anticipate a more comprehensive and accurate prediction of resistance to FOLFOX, providing valuable advances in the classification of CRC patients resistant or sensitive to FOLFOX. The importance of employing a successful DL classification model is crucial for precisely categorising patients into these resistance/sensitivity groups, facilitating precise treatment strategies based on robust predictive capabilities.

1.3.2 Thesis Summary

The work conducted in this thesis includes an extensive preprocessing pipeline designed to address background noise and artefacts in WSI effectively. This preprocessing workflow comprises several crucial steps, such as colour normalisation, stain separation, and tissue segmentation, aimed at standardising colour representations, isolating tissue stains, and delineating regions of interest. Additionally, quality control measures were implemented to detect and mitigate common artefacts, ensuring that the WSI data underwent thorough quality enhancement. This comprehensive preprocessing strategy played a pivotal role in optimising the accuracy and reliability of downstream analyses, ultimately facilitating robust and meaningful interpretations of the results. Furthermore, gene expression data underwent preprocessing, including deg, co-expression analysis, and hub gene identification within cancer gene expression datasets, aiming at dimensionality reduction. By pinpointing key genes that played pivotal roles in cancer biology, this process facilitated the reduction of dataset complexity while retaining essential information. The identified hub genes served as valuable biomarkers for cancer diagnosis, prognosis, and therapeutic target identification. Integration of this hub gene identification process into the overall framework for phenotype prediction enhanced the accuracy and interpretability of the predictive models, ultimately contributing to improved patient stratification and personalised treatment strategies in oncology.

1.4 Aims and Objectives

1.4.1 Aims

This study aims to build a predictive framework to classify CRC patients resistant or sensitive to FOLFOX. The scope of the research extends to include additional phenotypic tasks, specifically the prediction of the impact of cancer on healthy individuals and cancer patients, using DL methodologies from two distinct modalities including WSIs and PIs.

1.4.2 Objectives

1. Investigate the underlying mechanisms that contribute to resistance to FOLFOX by identifying and annotating key hub genes through a comprehensive analysis of high-throughput gene expression data obtained from CRC patient samples.

2. Develop a robust feature space using TDA on hub gene expression data for training the MLP model to predict the resistance of FOLFOX using topological summaries in the form of PIs.

3. Train a CNN model for predicting cancer stage based on pre-processed WSIs by implementing data preprocessing and model performance evaluation.

4. Enhance FOLFOX resistance prediction by integrating the strengths of trained CNN and MLP models through fusion methodologies.

5. Evaluate and compare the effectiveness of the developed framework with that of an established phenotype prediction tool, focussing on bimodalities.

Chapter 2

Materials and Methods

The purpose of this study was to develop a DL model to predict resistance to FOLFOX in patients with CRC using bimodal data, specifically PIs encoding gene expression data and WSI. The set of genes were pre-selected with a twofold objective: first, to enhance the understanding of resistance to FOLFOX through gene expression data and second, to streamline subsequent computational processes. This pre-selection process involved the identification of DEGs, followed by the extraction of hub genes using co-expression analysis. The utilisation of hub genes in the pre-selection of the gene expression matrix serves as a strategic approach to focus subsequent analyses on the most relevant molecular features. Subsequently, these specifically identified hub genes were encoded into shape using TDA to derive robust features.

The term pre-select denotes the preliminary steps of processing or selecting pertinent elements before engaging in downstream tasks, particularly within the context of topological studies. This term underscores the importance of carefully filtering data to ensure that only relevant information is processed for subsequent analysis. The over-representation analysis of the hub genes was used to identify biological functions related to resistance to FOLFOX. In the second mode of data, the WSIs were processed to make it suitable for input into a CNN. The developed framework in the present study, termed *GeTopology*, was compared with a previously published method called "non-small-cell lung cancer classification" (*NSCLC*) [199]. The prediction framework *GeTopology* was evaluated on four additional cancer datasets for the classification task of distinguishing between healthy individuals and cancer patients to further test its robustness and generalisability. The summarised framework is depicted in Figure 2.1.

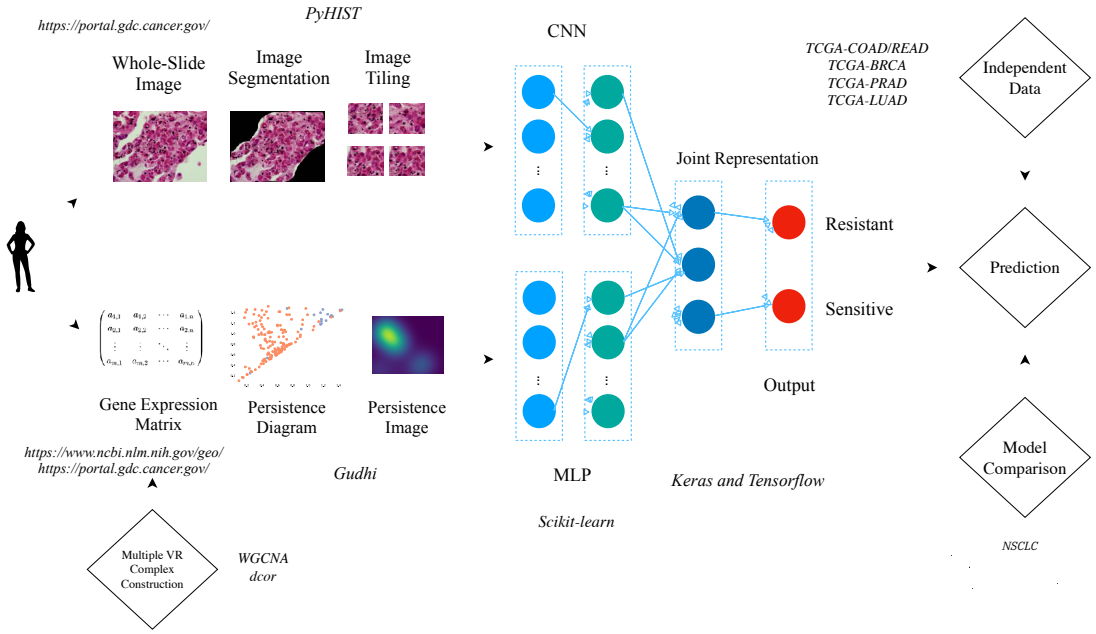


Fig. 2.1 The overall task of classifying CRC patients as resistant or sensitive to FOLFOX involves the use of persistent images (PIs) derived from gene expression data encoded in persistence diagrams (PDs). Additionally, the framework includes the use of whole slide images (WSI), subjected to image segmentation and tiling processes. Processed WSIs are used to train a convolutional neural network (CNN), while PIs are used to train a multilayer perceptron (MLP). The CNN and MLP are integrated to collectively classify CRC patients as either resistant or sensitive to FOLFOX. Further analyses encompass the construction of multiple Vietoris-rips (VR) complexes using various correlation metrics. The prediction framework is also validated using independent cancer datasets, including The Cancer Genome Atlas (TCGA) datasets for breast adenocarcinoma (BRCA), prostate adenocarcinoma (PRAD), lung adenocarcinoma (LUAD), and colon adenocarcinoma/rectal adenocarcinoma (COAD/READ). The non-small-cell lung cancer classification (*NSCLC*) tool, which takes gene expression and WSI data as input, is compared to the developed framework. Italicised words indicate the computational tools used for each task: Weighted correlation network analysis (*WGCNA*), *NSCLC*, distance correlation (*dcor*), and histological image segmentation tool (*PyHIST*). Data sources shown are from the National Center for Biotechnology Information Gene Expression Omnibus (NCBI-GEO) and Genomic Data Commons (GDC).

2.1 Medical Image Data

2.1.1 Data Collection

Data, including RNA-Seq count and WSI, were obtained from two studies conducted by The National Cancer Institute (NCI) and the Exceptional Responder Initiative (ERI) (https://portal.gdc.cancer.gov/projects/EXCEPTIONAL_RESPONDERS-ER) as well as the Human Cancer Model Initiative and the Cancer Model Development Center (HCMI-CMDC) (https://portal.gdc.cancer.gov/projects/HCMI-CMDC). The NCI was the primary factor in the selection of these data sources due to its extensive research throughout the United States, covering all demographics and its status as one of the leading cancer research organisations in the world [200]. Additional WSI data were obtained from The Cancer Genome Atlas (TCGA) (available at https://www.cancer.gov/ccg/research/genome-sequencing/tcga) database from the colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) cohorts to expand the number of WSI representing various stages of cancer. The published work from this dataset is summarised in Table 2.1. The NCI and the National Human Genome Research Institute (NHGRI) collaborated in 2006 to launch the TCGA programme. This collaborative effort has seen researchers from a variety of disciplines and organisations come together to study more than 20,000 primary cancer samples and related normal samples from 33 different types of cancer [201]. From the NCI database, the WSI data was accessed and downloaded using the Genomic Data Commons (GDC) Data Portal accessed from https://portal.gdc.cancer.gov for each CRC patient who was resistant or sensitive to FOLFOX. WSIs were classified according to cancer stage. A limitation of this dataset was the class imbalance, which was addressed by WSI data processing.

Table 2.1 The overview of published work from the ERI (Exceptional Responders Initiative) and HCMI-CMDC (Human Cancer Models Initiative - Colorectal Cancer Model Development Consortium) projects, as well as The Cancer Genome Atlas (TCGA) from the colon and rectal adenocarcinoma cohorts (COAD/READ) along with relevant references. The table lists the projects and associated publications, highlighting key contributions from each initiative. These references summarise the ERI and HCMI-CMDC project, covering both pilot and review studies on anticancer drugs, clinical participant information, and impactful research findings. Notable outcomes include the implication of mycobiota in gastrointestinal and lung cancers, suggesting that tumour-associated fungal DNA is a potential biomarker. Other studies focus on aligning tumour and cell line transcriptional profiles, using a developed tool, *Celligner* to enhance clinical translation, identifying autophagy-related lncRNAs in TCGA-COAD, and predicting colon cancer prognosis through machine learning analysis of transcription factor expression.

Project	**References**
ERI	[202–204]
HCMI-CMDC	[205, 206]
TCGA-COAD/READ	[207, 208]

2.1.2 Data Processing

In the application of WSIs to CNNs, a Canny edge detector was used to delineate the boundary of the segmentation mask and eliminate background noise [209]. This technique leverages the calculus of variations to diminish noise and identify the maximum and minimum of a given functional by making small changes in the functions [210], where functional refers to a function whose input is a function itself.

Gaussian Kernel and Canny Edge Detection

The Gaussian kernel, a convolutional filter, was used to apply a weighted average across neighbouring pixels, emphasising or de-emphasising certain features. A larger Gaussian kernel size contributed to reducing sensitivity to noise, smoothing out variations, and minimising the impact of individual noisy pixels. However, larger kernel sizes led to a slight increase in edge detection localisation error, necessitating a balance between noise reduction and edge localisation accuracy [211]. In this context, a kernel size of 5×5 was utilised, with slight adjustments possibly needed based on specific circumstances [212]. The Canny algorithm employed four filters to detect horizontal, vertical, and diagonal edges in the blurred image [212].

Gradient Computation

In the subsequent step, the gradient computation was performed using edge detection operators. These operators calculated the gradient of the image, revealing both the intensity and the direction of the edges [213]. Generally, filters that emphasised intensity changes in both horizontal (x) and vertical (y) directions were applied. After smoothing the image, the partial derivatives I_x and I_y with respect to x and y were computed [214]. Subsequently, the intensity of G was determined using Equation 2.1.

$$|G| = \sqrt{I_x^2 + I_y^2} \tag{2.1}$$

The slope θ of the gradient was calculated in Equation 2.2:

$$\theta_{x,y} = \arctan\left(\frac{I_y}{I_x}\right) \tag{2.2}$$

This resulted in the formation of contours that distinguish the outer lines of the object of interest.

Non-Maximum Suppression

Following this, a third algorithm was employed using non-maximum suppression to enhance the contrast between the background and the mask. Essentially, the algorithm traversed all points in the gradient intensity matrix, identifying pixels with the highest values in the edge directions. This process highlighted and sharpened the edges of the image, improving the distinction between the background and the mask [215].

Double Thresholding

A fourth algorithm was employed, which involves the application of a double-thresholding technique to categorise pixels into three types: strong, weak, and non-essential. Strong pixels had intensities high enough to unequivocally contribute to edge detection. Weak pixels had intensities that fell between high and low thresholds, and non-essential pixels were ignored in the edge detection process [216]. A high double threshold of 0.95 was set to identify strong pixels, while a low threshold identified

non-essential pixels. Pixels with intensities between the two thresholds were classified as weak. The subsequent hysteresis mechanism determined which weak pixels could be considered strong, achieving a binary intensity representation (strong and weak pixels) in the resulting threshold image [217].

Hysteresis for Edge Tracking

Finally, hysteresis, a technique in image processing where weak pixels are converted into strong pixels if they are adjacent to at least one strong pixel, was used to track edges. To mitigate false detections due to noise, the algorithm applied a Gaussian filter kernel to smooth the image, reducing the influence of noise on the edge detector. This step was crucial in preventing CNNs from learning false features during model training. Subsequently, a graph-based image segmentation technique was applied to assess the quality of the segmentation mask and background regions [210].

The format of the WSIs was converted from slide virtual slide (*SVS*) to portable network graphic (*PNG*) to reduce image size, improving downstream computational performance. Lastly, the WSIs were tiled to increase the training sample set and address class imbalance. The WSI processing steps mentioned above were executed using the Python package histological image segmentation tool (*PyHIST*) (version 0.36.0), available at https://pyhist.readthedocs.io/en/latest/ [218]. Following the tiling process (outlined in Figure 2.2), the count of WSIs is documented in Table 2.2. WSIs corresponding to patients with CRC resistant or sensitive to FOLFOX in the RNA-Seq dataset were excluded from tiling, as they were used solely to test the DL model using bimodalities.

Table 2.2 The number of tiled WSIs categorised by dataset and cancer stage. The table provides a breakdown of the tiled whole-slide images across different cancer stages within each dataset. Specifically, it outlines the number of tiled images corresponding to Stage I, Stage II, Stage III, and Stage IV in the categories of Folinic acid, Fluorouracil and Oxaliplatin (FOLFOX) resistance, FOLFOX Sensitive, and The Cancer Genome Atlas for colon and rectal adenocarcinomas (TCGA-COAD/READ).

Dataset	**Stage I**	**Stage II**	**Stage III**	**Stage IV**
FOLFOX Resistance	1	0	2	7
FOLFOX Sensitive	6	7	3	1
TCGA-COAD/READ	442	489	621	501

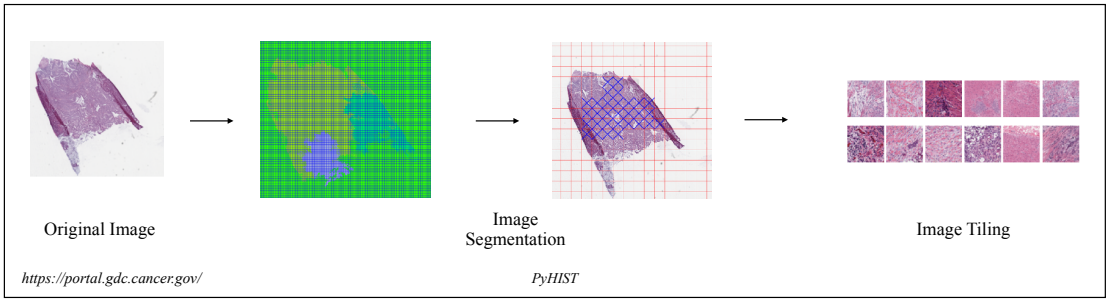


Fig. 2.2 Processing WSI including the separation of the background from the tissue of interest and the subsequent subdivision of the image into smaller sections through a tiling process. These steps are important for enhancing computational efficiency and preparing the WSI for further analyses and downstream tasks. The italicised words indicates the data source (i.e. the GDC - Genomic Data Commons data portal) and the Python programming package histological image segmentation tool - *PyHist* used to perform the processing of WSIs.

2.2 Gene Expression Data

2.2.1 Data Collection

As stated in Section 2.1.1, RNA-Seq count data were obtained from ERI and HCMI-CMDC. These data were normalised in terms of TPM. The sum of all TPMs for each sample is the same, making this normalisation technique suitable for comparative analysis between samples [111]. The number of patients with CRC resistant or sensitive to FOLFOX in the study was obtained from https://portal.gdc.cancer.gov/projects/EXCEPTIONAL_RESPONDERS-ER and https://portal.gdc.cancer.gov/projects/HCMI-CMDC. One notable research outputs from these datasets include, Conley *et al.* (2020) aimed to identify molecular features associated with a favourable response or outcome in various cancers, including CRC patients treated with FOLFOX [202]. Another study by Bilusic *et al.* (2021) found that genomic instability, along with treatment with DNA damage compounds and favourable anticancer immunity, may contribute to exceptional responses to standard anticancer therapies including FOLFOX [219]. Lastly, Wang *et al.* (2022) focussed on applying xenografts initiated from patient-derived organoids as a dependable cancer pre-clinical model in basic cancer research and precision medicine [220]. The sample numbers of the dataset are

summarised in Table 2.3.

Table 2.3 The set of RNA-Seq count data sample used in the study, downloaded from ERI and HCMI-CMDC. The table provides information on the number of patients included in each dataset, categorised as Folinic acid, Fluorouracil and Oxaliplatin (FOLFOX) resistance (10 patients) or FOLFOX sensitive (17 patients).

Dataset	**Number of Patients**
FOLFOX Resistance	10
FOLFOX Sensitive	17

For more than two decades, microarray technology was the mainstay method for quantifying gene expression, resulting in a higher number of public microarray datasets than RNA-Seq data. Raw probe intensity data was obtained from the Gene Expression Omnibus of the National Centre for Biotechnology Information (NCBI-GEO) repository (available at https://www.ncbi.nlm.nih.gov/geo/). The NCBI-GEO database could be accessed using three distinct entities: the sample (GSM), the series (GSE), and the platform (GPL). Each of these entities was assigned a unique accession number that could be used to directly access the data. The NCBI-GEO website was used to search and download public CRC patients resistant or sensitive to FOLFOX. The GSE accession numbers of the microarray samples obtained included GSE69657, GSE72968, and GSE81653 [221–223]. The sample numbers for each dataset are summarised in Table 2.4. The GSE69657 and GSE72968 datasets were generated using the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array with the GPL570 platform, while the GSE81653 dataset was generated using the Affymetrix Human Gene 2.0 ST Array with the GPL16686 platform. Patients in the GSE69657, GSE72968, and GSE81653 datasets received four cycles of FOLFOX chemotherapy. After evaluating the effectiveness of chemotherapy according to the Response Evaluation Criteria in Solid Tumours (RECIST), samples were taken [224]. These datasets were chosen because they were obtained from patients with CRC who had been treated with FOLFOX. These data provided a report of the stage of cancer and recorded the effects of FOLFOX treatment, including resistance or sensitivity. Ideally, more RNA-Seq data would have been used, given its fewer technical biases. However, due to data availability constraints, these microarray datasets were chosen.

For the MLP prediction task exclusively, the microarray datasets were merged into a single gene expression dataset to generate training data which were later encoded into topological summaries, where the number of training samples was 82 (CRC patients

Table 2.4 The microarray data sample set used in the study comprises samples resistant and sensitive to Folinic acid, Fluorouracil, and Oxaliplatin (FOLFOX). The following table summarises the datasets along with the number of samples for each category and includes the corresponding uniform resource locators (URLs) for access. Furthermore, a multilayer perceptron model (MLP) is mentioned, *MLP Training* dataset with 82 resistant and 115 sensitive samples used for MLP training. Additionally, references are provided for further information. Notable research outputs from provided references include the identification of myeloid ecotropic viral integration site 1 (MEIS1), a transcription factor down-regulated in colorectal cancer, which promotes tumorigenesis and resistance to oxaliplatin. Furthermore, precision medicine-guided co-delivery of a small interfering RNA (siRNA) targeting the Asporin (*ASPN*) gene and oxaliplatin overcomes chemoresistance, while chemotherapy suppresses Sonic Hedgehog (*SHH*) gene expression. Finally, a gene expression signature predicts the response to chemotherapy in CRC patients, validated by systems biology analysis and experimental transcriptomics data validation.

Dataset	**Resistant to FOLFOX**	**Sensitive to FOLFOX**	**URL**	**References**
GSE69657	9	6	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69657	[225–227]
GSE72968	9	17	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72968	[228]
GSE81653	64	102	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81653	[229–231]
MLP Training	82	115		

resistant to FOLFOX) and 115 (CRC patients sensitive to FOLFOX). The merging of samples across different datasets was done to increase the number of training samples. However, datasets were individually analysed for hub gene pre-selection to relate them to the biological function. The focus on hub genes had two major advantages for downstream tasks including, selection of driver genes related to FOLFOX resistance and reduction of computational cost for topology-based experiments. By focussing on the hub genes of patients resistant to FOLFOX, the analysis identified genes that play an important role in resistance to FOLFOX treatment. This focus approach allowed for more targeted research of genetic factors specifically related to resistance. Moreover,

by excluding FOLFOX-sensitive groups, computational resources could be preserved for downstream topological tasks, and analysis could be simplified to focus on the potential for greater influence of the hub genes of the resistance group.

2.2.2 Data Processing

The RNA-Seq data were normalised by calculating the mean TPM count for each gene in all samples. The TPM count for each gene in each sample was then divided by the mean TPM count. The median number of transformed gene counts for each sample was used to determine the size factor for that sample, which was used to adjust for the size of the library and the bias in the composition of RNA [111]. Outliers were identified and excluded from the analysis using Cook's distance. The Cook's distance is a statistical measure used in regression analysis to detect data points that could have a major effect on the estimation of regression coefficients [232]. It evaluates the effect of eliminating a certain data point from the analysis by calculating how much the regression coefficients vary. Greater Cook's distance values suggest that taking out a certain observation would significantly modify the regression model [233]. By recognising and excluding outliers based on Cook's distance, analysts can improve the robustness and accuracy of regression analyses [234].

Furthermore, genes with a mean normalised count below a threshold of ten, determined by an optimisation process were removed, which improved the detection power by reducing the severity of multiple-test adjustment of the p-value during differential gene expression analysis [235–237]. Gene removal was performed using the R programming package *DESeq2* (version 1.42.0) [131].

For the raw microarray dataset, RMA was used to perform background correction, normalisation, summation, and quantification of gene expression. Smooth quantile normalisation was used to adjust the data in a less drastic way and to achieve a balance between variance and bias. This technique was also used to avoid oversimplification of the data and was implemented using the R programming tool *vsn* (version 3.70.0) [238]. The microarray intensity values for the protein-coding probeset were determined by summarisation, eliminating any unnecessary probes from the array. To perform RMA, *affy* (version 1.80.0) was used [85]. The microarray probes were then matched to the gene symbols using the Ensembl gene annotation database (available

at https://www.ensembl.org/info/data/biomart/index.html) [239].

2.3 Gene Set Pre-selection

Figure 2.3 describes the process of recognising the hub genes from gene expression data, which is composed of two main stages. Firstly, DEGs are identified and their biological functions are carefully described. Secondly, hub genes are identified and described through co-expression analysis. It is important to note that the identified hub genes are essential for pre-selecting gene expression data, thus forming the basis for further analyses. This two-phase process is necessary to uncover the complex molecular landscape and potential biomarkers within the gene expression data.

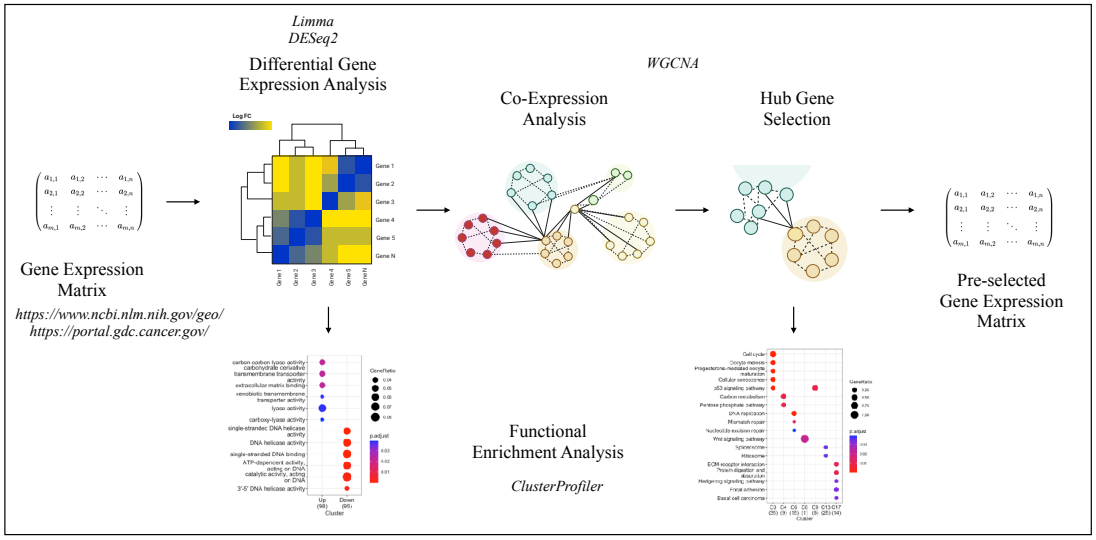


Fig. 2.3 The process to identify hub genes from gene expression data involved two steps. First, differential gene expression analysis was performed to identify DEGs, and their biological functions were annotated. Then, hub genes were identified and annotated by performing co-expression analysis. Hub genes were used to pre-select gene expression data. Italicised words indicate the computational packages used for the specified task, including *limma*, *Deseq2*, *clusterProfiler* and Weighted Gene Correlation Network Analysis (*WGNCA*). In addition the data sources from the National Center for Biotechnology Information - Gene Expression Omnibus (NCBI-GEO) and Genomic Data Commons (GDC) were used.

2.3.1 Differential Gene Expression Analysis

The initial gene set of 19,116 protein-coding genes (the number of known nuclear protein-coding genes) was reduced through pre-selection to gain insight into the resistance to FOLFOX from gene expression data and to enhance the efficiency of subsequent computational processes [240]. This led to a decrease in the dimensionality of the gene expression dataset. The reduction occurred in two stages, involving the assessment of both differential gene expression and co-expression. Differential gene expression analysis was conducted to compare different sample types and identify genes regulated in CRC patients resistant or sensitive to FOLFOX treatment. For this analysis, the R programming packages *limma* (version 3.58.1) [83] (for microarray data) and *DESeq2* (version 1.42.0) (for RNA-Seq data) were employed.

limma

limma fitted a linear regression model for all pairwise gene comparisons among samples, as depicted in Equation 2.3.

$$\mathrm{y_g} = \mathrm{X} \times \delta_{\mathrm{g}} \tag{2.3}$$

The variable $\mathrm{y_g}$ was estimated by fitting a linear regression model. The coefficient δ_g of interest could then be derived from the gene expression values (X) for each gene. The empirical Bayes technique (eBayes) was used to moderate overall gene-wise variances by shrinking the variance [241]. The average variance trends of the global inter-sample covariance were calculated to determine a weighting system to rank all gene-gene correlations. The weight function reduced the significance of gene-to-gene correlations that had a high covariance. This method reduced the level of noise by taking into account extreme values and strong correlation, thus decreasing the number of false positives (FP) [242].

The coefficient δ_g was used to test the gene sets and prioritise genes. For each predicted δ_g, a t-test was used to compare the mean of a pairwise comparison between the FOLFOX resistant group and the sensitive samples. The Benjamini-Hochberg probability value (BH-adjusted p-value) was used to evaluate the statistical test for each gene and adjust the p-value to reduce the false discovery rate [243]. Genes with a

BH-adjusted p-value less than 5% were considered statistically significant. Genes with a logFC greater than 0.5 were identified as up-regulated, while those with a logFC of less than -0.5 were determined to be down-regulated. A logFC of 0.5 was used to increase the number of identified DEGs. Although this threshold is low, this change in FC can provide information on the magnitude of the expression difference [244]. It is essential to note that even slight modifications in gene expression can have a considerable biological impact [91]. DEGs were employed in the pre-selection of gene expression data before conducting co-expression analysis. FC is essential to provide a more meaningful understanding of the distinctions between patients with CRC who are resistant or sensitive to FOLFOX.

DESeq2

DESeq2 fitted a linear regression on the expression of each gene to follow a negative binomial distribution with and dispersion α_i. The mean $\mu_{i,j}$, was taken as a quantity $\mathrm{q}_{i,j}$, proportional to the concentration of gene cDNA fragments in the sample, scaled by a normalisation factor $\mathrm{s}_{i,j}$ summarised by Equation 2.4.

$$\mu_{i,j} = \mathrm{s}_{i,j}\, \mathrm{q}_{i,j} \tag{2.4}$$

The design matrix elements x_j and its coefficients γ_i were used in the linear model to compare CRC patients resistant or sensitive to FOLFOX. The returned coefficients represented the overall expression of the gene and the logFC between the experimental groups and are shown by Equation 2.5.

$$\log_2 \mathrm{q}_{i,j} = \sum \mathrm{x}_j\, \gamma_i \tag{2.5}$$

Similarly to the implementation in *limma*, a linear model was used, and subsequently, the gene-wise dispersion estimates were adjusted based on the predicted values to derive the final dispersion values. The eBayes function determined the amount of shrinkage based on the discrepancy between the predicted and actual gene-wise dispersion values [241]. A Wald test was used for the significance test to identify DEGs. The Wald

test, a widely used test in bioinformatics, provided robust results in the context of gene expression analysis. Particularly suitable for detecting differential expression by assessing the significance of coefficients in the context of a linear model, the Wald test emerged as a reliable choice for this analytical task [245]. For each statistical test, the adjusted p-value of BH was calculated. Genes with a p-value lower than 5% after BH correction were identified as significantly altered. Similarly, genes with a logFC greater than 0.5 were identified as up-regulated, while those with a logFC of less than -0.5 were determined to be down-regulated. Significant DEGs were then used to pre-select gene expression data prior to co-expression analysis.

2.3.2 Co-expression Analysis

Co-expression analysis is useful for the interpretation of high-dimensional gene expression data [246]. Co-expressed gene modules can be identified to uncover related biological processes in patients resistant to FOLFOX treatment. The study specifically investigated FOLFOX-resistant CRC patients, focussing exclusively on resistance dynamics. Furthermore, the FOLFOX resistance framework developed was fine-tuned to the complexities of resistance mechanisms. By using weighted co-expressed networks, the overarching goal was to identify hub genes and their biological functional linkages. Furthermore, identified hub genes were used to further pre-select gene expression data.

Co-expressed gene networks are based on correlation, which is a statistical measure of the relationship between two random variables. Pearson's correlation is the most commonly used to build these networks, since they focus on capturing the linear relationship between the variables [247]. An alternative to Pearson's correlation measures is the biweight mid-correlation (BiCor), which is becoming increasingly popular due to its resistance to outliers and its ability to handle normality distortions [248]. Gene expression data was transformed by calculating the BiCor across all gene pairs. A correlation coefficient close to one indicates that a gene pair is strongly related, while a coefficient close to zero implies a weak relationship [249]. The BiCor was determined using Equation 2.6.

$$\text{bicor}\,(\mathrm{x}, \mathrm{y}) = \sum_{i=1}^{m} \tilde{\mathrm{x}}_i \tilde{\mathrm{y}}_i \tag{2.6}$$

The BiCor equation was computed using two vectors, x and y. These vectors were made up of gene pairs x_1, x_2 ..., x_m and y_1, y_2, ..., y_m in patients from gene expression data. To compensate for outliers, the weight of each gene expression measure was determined by its proximity to the median gene expression value. A higher weight was assigned to gene expression values closer to the median gene expression value. The total sum of the weights was indicated by i for each gene pair calculation. This helped to reduce the impact of outliers and adjusted for extreme differences between gene pair comparisons [248]. After calculating BiCor, a weighted adjacency matrix was calculated as shown in Equation 2.7.

$$\mathrm{a}_{i,j} = \mathrm{bicor}_{i,j}^{\beta} \tag{2.7}$$

A weighted network was constructed using soft thresholding (β) of the correlation coefficient. This preserves continuous correlation data, leading to highly reliable results [250]. In comparison, the results of unweighted networks, constructed by thresholding a pairwise association measure, are often highly dependent on the threshold [143]. The weighted adjacency for the i^{th} and j^{th} gene ($\mathrm{a}_{i,j}$) was calculated by soft thresholding the BiCor ($\mathrm{bicor}_{i,j}$). The smallest value of β was determined based on the scale-free topology criteria, which necessitated its selection to closely approximate a scale-free network structure. A scale-free network follows a power-law distribution that exhibits metabolic, protein, and gene interaction networks based on the distribution of the number of node connections in the network.

The weighted adjacency matrix obtained from the construction of the gene network was transformed into a signed-topological overlap matrix (signed-TOM). This transformation is crucial as TOM methods offer increased precision in determining the interconnectivity between gene pairs, thereby reducing the susceptibility to false connections or connections influenced by random noise [144, 143, 251]. The signed-TOM specifically measured the strength of connections between neighbouring genes with a direct relationship and also accounted for non-direct gene associations. In contrast, unsigned-TOM treats gene pairs with negative correlations as unconnected [251]. The transformation of adjacency values into signed-TOM was executed using Equation 2.8.

$$\mathrm{w}_i, j = \frac{|\,\mathrm{l}_{i,j}, + \mathrm{a}_{i,j}\,|}{min(\mathrm{k}_i, \mathrm{k}_j) + 1 - \mathrm{a}_{i,j}} \tag{2.8}$$

The signed-TOM measure, $\mathrm{w}_{i,j}$, is calculated for the i^{th} and j^{th} genes. The number of genes connected to genes i and j is indicated by $l_{i,j}$. The weighted adjacency measures are denoted by $\mathrm{a}_{i,j}$. The variables k_i and k_j indicate the measure of gene interconnectivity between genes i and j.

The computed signed-TOMs were grouped using hierarchical grouping of average links to reveal gene modules. The average distance between the nodes in the gene modules was calculated. This was followed by hierarchical clustering, which involved recursively splitting the dataset into gene modules [252]. Following previously published literature, the maximum height threshold of the dendrogram was adjusted to 0.25 (indicating that the correlation threshold is 0.75), to modulate the size and number of gene modules identified and to test the sensitivity of the clustering technique used [253–259]. This prevented the identification of too many (more than 15) or too few (less than 2) gene modules. After hierarchical clustering of signed-TOM networks, the module membership for each gene was determined using Equation 2.9.

$$\mathrm{K}_{\mathrm{cor},i} = \mathrm{cor}(\mathrm{x}_i, + \mathrm{E}_q) \tag{2.9}$$

The module membership of node i in module q was represented by $\mathrm{K}_{\mathrm{cor},i}$. The correlation between the expression level of node x_i and the intramodular connectivity was indicated by E_q. Intramodular connectivity, also known as eigengene-based connectivity (k_{ME}), was used as a measure of the correlation between each gene expression value and the first principal component (PC) of a given module. The first PC served as a single representative measure of the gene expression value in a gene module [144]. To identify the hub genes, varying thresholds (ranging from 0.6-0.95) were set on the calculated k_{ME} values to select hub genes in each gene module (Section A). Hub genes, characterised by their connection to many other genes in the module, were identified by ranking those with the highest k_{ME} values. Hub genes were considered the main regulators of the gene modules that influence biological behaviour [145]. Gene expression matrices were pre-selected based on identified hub genes, and over-representation analysis was performed to explore their biological functions. All the steps mentioned above were

carried out using the R programming package Weighted Correlation Network Analysis (*WGCNA*) (version 1.72-1) [144, 143].

2.3.3 Over-representation Analysis

Over-representation analysis methods were applicable to both transcriptomic data from microarrays and RNA-seq. Both DEGs (Section 2.3.1) and hub genes were related to biological functions by performing over-representation analysis from the biological pathway database, KEGG [260], and the ontology database, GO [260]. The purpose of this task was to discover the hub genes that were used to create a feature space for topological summaries. This initial step was necessary to give more meaning to the topological features. To minimise the redundancy of biological functions, a hypergeometric distribution was used to assess if a certain gene set and its associated biological terms were over-represented. The hypergeometric p-value was determined by computing the probability of randomly selecting k or more successes from a population of n draws [261]. In other words, let variable A represent a GO term or the set of genes associated with it. Let S be a set of genes that satisfy a specific criterion (e.g., DEGs or hub genes) from a comprehensive gene list G in an experimental context, which included 19,116 protein-coding genes. The number of genes present in sets S and A indicated the presence of A in S. Assuming that S and A are independent (that is, the GO term is not related to the gene cluster), both S and A followed a hypergeometric distribution [262]. The p-value that measured the significance of the association was the probability of observing S and A [263]. The p-value was then corrected for with BH to reduce the FDR. Biological terms that had a BH p-value less than 5% were considered statistically significant and reported accordingly. Over-representation analysis was performed using the R programming package *clusterProfiler* (version 3.17) [162].

2.4 Data Correlation Analysis

2.4.1 Distance Correlations Measures

Before computing the correlation, the combined patients of the microarray datasets were divided into subsets: 70% training (M_{train}) and 30% testing (M_{test}) by random shuffling. This helps prevent any unintentional bias or patterns in the data order that could impact the model's performance during training and evaluation. In ML/DL

and statistical modelling, the ratio of 70 to 30 is common. A training set (70%) is used to train and develop models to learn data patterns and relationships. The remaining 30% of the test sets are independent datasets that the model did not see during training [264, 265]. This division was achieved through Monte Carlo cross-validation, a technique that involves randomly partitioning the dataset into multiple subsets for both training and testing. Monte Carlo cross-validation helps assess the performance and generalisability of the model by repeatedly resampling and evaluating its performance in different subsets of the data [266]. This process was repeated five times for each experiment. However, patients with gene expression generated from RNA-Seq were not split but still underwent a correlation transformation and were subsequently used as M_{test} due to their low sample counts. Moreover, the patients were split in this regard to prevent data leakage in subsequent DL tasks. Data leakage refers to the intentional introduction of data from testing or validation into the training process, which can lead to an overly optimistic estimate of the performance of the model [267]. A distance correlation matrix was calculated from the training (M_{train}) subset. As such, downstream DL tasks were trained exclusively on training data.

Distance correlation is a measure of dependence between two sets of random variables that may not necessarily have the same dimensions. It is different from Pearson's correlation, which can only detect linear associations between two random variables and assumed normality [268]. Distance correlation, on the other hand, was able to measure linear and non-linear associations between two random variables or vectors [269]. This made it particularly useful in biological networks, where it could be used to capture complex relationships between genes [270]. The distance correlation was derived from other quantities, including distance covariance, variance, and standard deviation shown in Equations 2.10, 2.11 and 2.12, respectively.

$$\mathrm{dCov}^2(I, J) = \mathrm{cov}(\|I - I'\|, \|J - J'\|) - 2\,\mathrm{cov}(\|I - I'\|, \|J - J''\|) \tag{2.10}$$

Pearson's covariance between the pair of random variables I and J was denoted by *cov*. The prime random variables *I'*, *I'* and *I"*, *J"* were independent and identically distributed copies of I and J, respectively. This meant that each of these random

variables had the same probability distribution and were mutually independent.

$$\mathrm{dVar}^2(I) := \mathrm{E}[\|I - I'\|^2] + \mathrm{E}^2[\|I - I'\|] - 2\,\mathrm{E}[\|I - I'\|\,\|I - I''\|] \tag{2.11}$$

The expected value of a random variable with a finite number of outcomes was the average of all possible outcomes, and each outcome was weighted according to its probability. Here, *I*, *I'*, and *I"* were independent and identically distributed random variables, and *E* was the expected value.

$$\mathrm{dCor}^2(I, J) = \frac{\mathrm{dCov}^2(I, J)}{\sqrt{\mathrm{dVar}^2(I)\ \mathrm{dVar}^2(J)}} \tag{2.12}$$

The distance correlation between two random variables was determined by dividing their distance covariance by the product of their distance standard deviations [271]. All pairwise distance correlation measures for genes in $\mathrm{M_{train}}$ were calculated using the Python programming package *dcor* (version 0.6) [270].

2.4.2 Signed-Topological Overlapping Measures

Similarly, prior to computing the signed-TOM correlation, the patients from the microarray datasets were divided into two subsets: a training set ($\mathrm{M_{train}}$) comprising 70% of the data and a testing set ($\mathrm{M_{test}}$) comprising the remaining 30% by random shuffling. However, patients with gene expression generated from RNA-Seq were not split but still underwent a correlation transformation and were subsequently used as $\mathrm{M_{test}}$. The idea of signed-TOM was presented to indicate the degree of interconnection between two nodes, which was not only determined by the direct connection between them but was also affected by indirect routes between them. Therefore, signed-TOM could help uncover indirect connections. In signed-TOM networks, nodes that had a negative correlation were not connected and the strength of their connection was zero or close to zero. The signed-TOM subtracted anti-reinforcing connections from the direct (and reinforcing) connections [272]. To achieve this, TOM required not only the connection strengths $a_{i,j}$ but also the indication of the correlations that were at the root. This could be conveniently accomplished by creating a modified adjacency

matrix $\widetilde{a}_{i,j}$ as demonstrated in Equation 2.13.

$$\widetilde{\mathrm{a}}_{i,j} = a_{i,j} \times sign(\mathrm{cor}(\mathrm{x}_i, \mathrm{x}_j)) \tag{2.13}$$

Where $\widetilde{a}_{i,j}$ is the modified adjacency matrix, $\mathrm{a}_{i,j}$ was the adjacency matrix calculated as shown in Equation 2.7. Variables x_i and x_j denoted the measure of correlation and direction (*sign*) of gene expression for pairs of genes. The signed-TOM was then defined as Equation 2.14.

$$\mathrm{w}_{i,j} = \frac{|\,\mathrm{l}_{i,j}, +\widetilde{\mathrm{a}}_{i,j}|}{min(\mathrm{k}_i, \mathrm{k}_j) + 1 - \widetilde{\mathrm{a}}_{i,j}} \tag{2.14}$$

Where $\mathrm{w}_{i,j}$ is the signed-TOM, k_i and k_j denoted the connectivities of nodes i and j. The number of genes connected to genes i and j is indicated by $\mathrm{l}_{i,j}$. The signed-TOM correlation for all gene pairs was calculated using the R programming package *WGCNA* (version 1.72-1) [144].

The distance correlation and signed-TOM were both useful methods for expressing the finite structure of a graph, as they succinctly illustrated the connection between the graph's nodes. As such, both metrics were used to represent the VR complex for the PH calculation.

2.5 Persistence Homology

2.5.1 Persistence Homology Computation

TDA was employed to analyse high-dimensional, incomplete, and noisy datasets, which could be challenging with traditional methods. TDA facilitated the detection of patterns and connections in the data that might not have been apparent with conventional techniques. The motivation for this research was to analyse the structure of gene expression data. Topology, a branch of mathematics, combined algebraic topology and other mathematical techniques to precisely examine the shape, with the main TDA tool being PH [273].

For each patient, PH was computed from M_{train} and M_{test} using persistence-based filtration (ϵ) on the input data described in Sections 2.4.1 and 2.4.2. During PH computation, the topological characteristics of up to three homology groups were identified, including *Betti*-0, *Betti*-1, and *Betti*-2, as explained in Section 1.2.4. Homology groups greater than three were too demanding to calculate using classical computers. Persistence-based filtration was performed up to the last edge between all graph nodes in the VR complex. To execute this operation, a weighted distance correlation/signed-TOM (G) was calculated using Equation 2.15 for each patient.

$$G_{i,j} = D_{i,j} \times (F_i + F_j) \tag{2.15}$$

The weight function is described by F_i and F_j, which modifies the correlation measure (distance correlation/signed-TOM) denoted by $D_{i,j}$. The weight functions represent the measures of gene expression measure (F_i and F_j) for the i^{th} and j^{th} gene for a single patient. Consequently, patients with higher gene expression measures resulted in higher $G_{i,j}$ values, while those with smaller gene expression measures had lower $G_{i,j}$. This led to each patient having a different correlation matrix, translating to a different VR complex for the PH calculation. PH calculations were performed and PDs were generated for each patient based on the signed-TOM and distance correlation metrics.

2.5.2 Vector-Based Transformation

To use PDs in DL models, vector-based transformations in the form of PIs were performed. PIs were resistant to continuous modifications, making them a great option for DL applications, as they could tolerate topological changes without losing stability [179]. PIs represented a smoothed version of the topological feature coordinates in the PD, created using a Gaussian kernel at each topological point [274]. The collection of pixels was arranged on a fixed rectangular grid. The PI of a given PD was a collection of pixels expressed in Equation 2.16.

$$R = [a, b] \times [c, d] \tag{2.16}$$

The area of each pixel in the image was $[a, b] \times [c, d]$, which translated the PD into pixels. The intensity of each pixel was based on the weighting function for each characteristic. The entire pixelated area of all PDs generated for all patients was 80×80. If the dimensions were larger, the resolution of the topological pixels would decrease, or if the dimensions were smaller, it would increase. This represented a trade-off to capture high resolution information without using too much data size. In regions of the PI where the topological characteristics were more significant, the intensity of the pixels was higher than in those where they were less important [274]. The magnitude of the intensity of the pixels was determined by the absolute value of Equation 2.17.

$$\mathrm{P}_i = |\, \mathrm{b}_i - \mathrm{d}_i \,| \tag{2.17}$$

Here, $\mathrm{P_{i,j}}$ represented the intensity of the pixel for the i^{th} topological feature, the topological characteristics of each point was defined by ($\mathrm{b_i}$, $\mathrm{d_i}$).

Subsequently, the pixels in the PI were converted into fixed-dimensional vectors and used for downstream DL tasks. The Python package *gudhi* (version 2.0.0) was used to calculate PH and transform PI vectors [275]. The workflow is succinctly described in Figure 2.4. The processed WSI and the PIs for each patient were used for image classification and topological summary classification, respectively. Following the encoding of gene expression data into topological summaries in the form of PIs, CNNs were trained using WSIs. The CNNs were initially trained to initialise the learned weights before deploying the DL model using bimodalities. The binary classification approach using CNNs aimed to predict cancer stage, specifically distinguishing between cancer stage I and cancer stage IV.

2.6 Deep Learning

2.6.1 Image Classification

A CNN in the form of a MobileNet was used to classify WSIs by cancer stage, as illustrated in Figure 2.5. The last hidden layer of the MobileNet model was trained using the tiled WSIs to increase the training set and address class imbalances between cancer

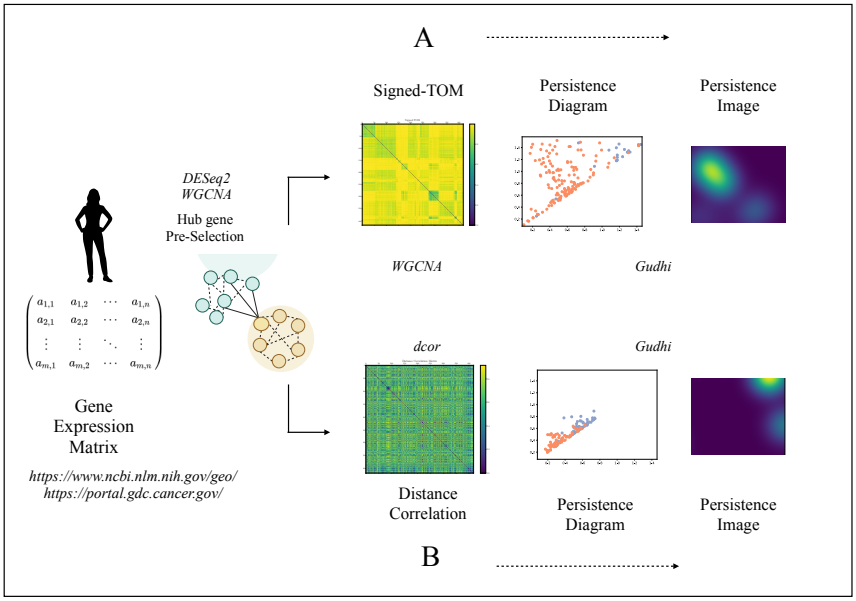


Fig. 2.4 Topological summaries derived from the VR complex were computed using both the distance correlation and signed-topological overlapping measures (signed-TOM) approaches. These summaries were then compared using persistence diagrams (PDs). Prior to phenotype prediction, vector-based transformations of the PDs into persistence images (PIs) were performed. The computational tools utilised include Weighted Gene Correlation Network Analysis (WGCNA), distance correlation (dcor), DESeq2, and gudhi. Additionally, the data sources from the National Center for Biotechnology Information-Gene Expression Omnibus (NCBI-GEO) and Genomic Data Commons (GDC) are shown.

stages I and IV, as described in Section 2.1.2. MobileNet, a form of transfer learning, was chosen for its advantageous characteristics. This model was pre-constructed, and its parameters having undergone fine-tuning. Transfer learning models, as demonstrated in previous studies, exhibit effectiveness even when confronted with a limited number of training samples [276–278]. The fundamental idea behind transfer learning was to take advantage of the knowledge acquired from a task with abundant labelled training data and apply it to a new task characterised by limited data [276]. Instead of starting the learning process from scratch, the model had already acquired patterns to solve a related task. The MobileNet model was selected for its lightweight structure, comprising 28 neural layers and a significantly smaller number of training parameters compared to other transfer learning models, such as the ResNet model [279]. The MobileNet was applied using the Python programming packages *keras* (version 2.10.0) and *Tensorflow* (version 2.9) [280, 281]. The MobileNet that had been trained was

then compressed and stored in a *pickle* module for use in a downstream bimodal DL task.

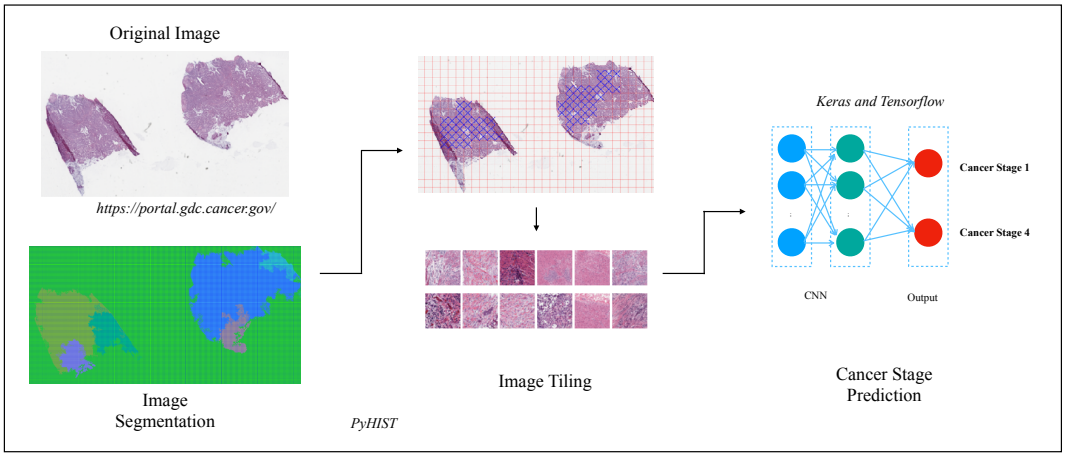


Fig. 2.5 The processing of WSI involved two key steps: image segmentation and image tiling. In the first step, image segmentation was applied to partition the WSI into meaningful regions. Following segmentation, the images are tiled, dividing them into smaller, manageable units. Subsequently, a convolutional neural network (CNN) employing the MobileNet architecture was utilised for prediction learning on the segmented and tiled images. This comprehensive approach enabled the extraction of relevant features from the WSI, facilitating effective image analysis and classification. Italicised words indicate computational tools used including the histological image segmentation tool (*PyHIST*) and data source from the Genomic Data Commons (GDC).

To perform WSI classification, the WSIs obtained from the TCGA-COAD/READ cohort were tiled and split into 70% training, 20% testing, and 10% validation sets by random shuffling. Due to computational limitations, the training was performed over 50 epochs and over five Monte Carlo cross-validation steps. The performance of the model was evaluated by calculating the mean accuracy, true positive rate (TPR), macro F1 score, and their respective errors. The accuracy of a model was determined by computing the number of correct classifications divided by the total number of classifications made, as defined by Equation 2.18:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2.18}$$

The proportion between true positives (TP) and true negatives (TN) relative to the sum of TP, TN, FP, and false negatives (FN) is known as the TPR or recall. The TPR was calculated using the following equation 2.19:

$$TPR = \frac{TP}{TP + FN} \tag{2.19}$$

The F1 score was derived from precision (Equation 2.20) and recall scores (Equation 2.19), representing the harmonic mean of the two. Contrary to the statement, the harmonic mean does not inherently encourage similar values for precision and recall. Instead, it emphasises the impact of lower values on its calculation, penalising extreme differences. Higher discrepancies between precision and recall result in lower F1 scores [282].

Regarding micro and macro F1 scores, these were aggregate metrics for multi-class classification. The micro F1 score was the overall F1 score calculated considering the total number of TP, FP, and FN in all classes [283]. In contrast, the macro-F1 score was the average of the F1 scores for each individual class. The micro F1 score was particularly sensitive to performance in larger classes, while the macro F1 score gave equal weight to each class, making it useful when class imbalances were present in the dataset [284, 285].

Recall was used to assess the prediction model's capability to accurately identify samples belonging to the positive class from the total number of samples in that class. It was calculated by dividing the number of TP by the sum of TP and FN [286]. Precision, on the other hand, measured the ratio of correctly identified positive results to the total number of positive results identified. This was mathematically expressed as the number of TP divided by the sum of TP and FP, equivalent to the TPR [287]. The precision was determined according to Equation 2.20.

$$Precision = \frac{TP}{TP + FP} \tag{2.20}$$

The F1 score determined for each predicted class (that is, resistant or sensitive to FOLFOX), defined as the harmonic mean of precision and recall, served as a measure

of the model's accuracy, calculated using Equation 2.21.

$$F1_i = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2.21}$$

The macro F1 score was determined by equation 2.22.

$$F1 = \frac{1}{n} \sum_{i=1}^{n} \mathrm{F1}_i \tag{2.22}$$

Where n is the number of classes and $\mathrm{F1}_i$ is the F1 score for class i according to equation 2.21.

The ROC curve evaluated the performance of a classification model by graphing the TPR (sensitivity) on the y-axis and the FPR (specificity) on the x-axis across different threshold values. This graphically illustrates the cumulative distribution function (CDF) of the detection probability on the y-axis and the CDF of the FP probability on the x-axis [288]. The area under the curve (AUC) served as a comprehensive metric, summarising the performance of the model across all potential classification thresholds [289].

2.6.2 Topological Summaries Classification

Patient PIs generated from the combined microarray datasets were used to train an MLP model, as shown in Figure 2.6. MLP models were simple neural networks that performed well for classification tasks. These models were appealing for handling complex and non-linear data, such as the relationship between topological features. Furthermore, MLPs were chosen because of their robustness to the amount of training data; even when the number of training data was low, they still performed well [290].

The PIs of the patients in the combined microarray dataset, constructed from $\mathrm{M}_{\mathrm{train}}$ were used to train the MLP model. Subsequently, PIs constructed from $\mathrm{M}_{\mathrm{test}}$ were used to test the trained MLP for the prediction of resistance to FOLFOX in patients with CRC. During model training, a feed-forward algorithm was used to classify patient PIs. Following hyperparameter tuning using the *GridSearchCV* function from *scikit-learn* (version 1.2.1) [291], the MLP was configured with ten neural layers, a

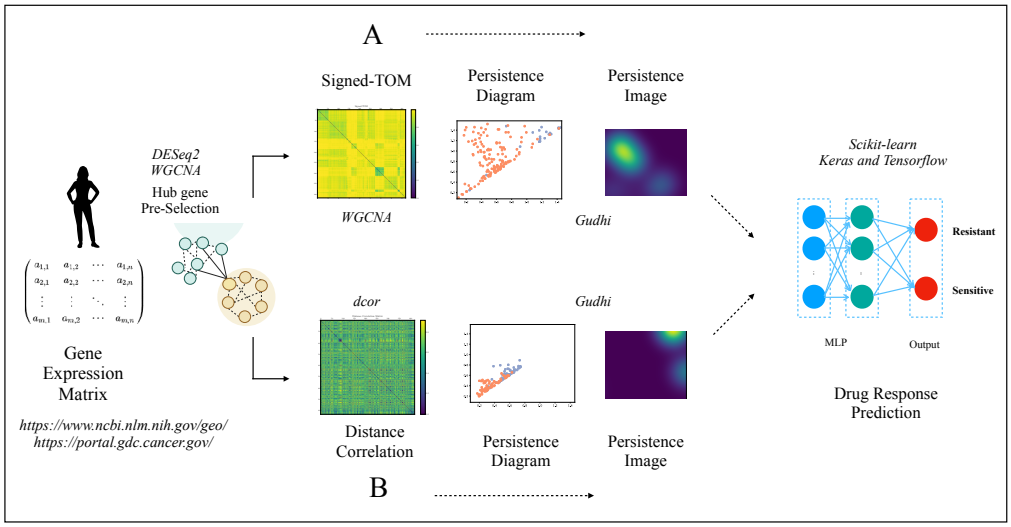


Fig. 2.6 Topological measures for predicting FOLFOX responses. Gene expression data was transformed into either signed-topological overlapping measures (signed-TOM) or distance correlation measures. Topological summaries, represented as persistence diagrams (PDs), were constructed for each patient and then vectorised into persistence images (PIs) before being input into a multilayer perceptron model (MLP) for the classification of FOLFOX resistance/sensitive CRC patients. Italicised words include computational tools including *DESeq2*, Weighted Gene Correlation Network Analysis (*WGCNA*), distance correlation (*dcor*) and *gudhi*. Furthermore, data sources from National Center for Biotechnology Information-Gene Expression Omnibus (NCBI-GEO) and Genomic Data Commons (GDC) are also shown.

ReLU activation function with regularisation steps in the hidden layers and a SoftMax activation function in the output layer. The MobileNet and MLP activation functions for the output layer were SoftMax, which returns a vector of interpretable probability scores for the predictions made [292]. These could then be combined to form a DL model using bimodalities. The neural weights were adjusted through 500 epochs of forward and backward propagation after model training. Monte Carlo cross-validation was carried out for five iterations and model performance was assessed using mean accuracy, TPR and macro F1 score, as well as their respective errors [293]. The MLP trained with PIs generated from M_{train} was compressed and stored (in a *pickle* module) and the classification task was performed using the Python programming packages *keras* (version 2.10.0), *Tensorflow* (version 2.9) and *scikit-learn* (version 1.2.1) [280, 281, 291]. After training a MobileNet with an expanded number of WSIs and

MLP using PIs (generated from signed-TOM and distance correlation metrics), the DL model using bimodalities was used to build a robust prediction modelling framework.

2.6.3 Classification using Deep Learning Model using Bimodalities

Cancer prognosis plays a crucial role in survival rates, with an estimated survival rate 90% in the early stages compared to a significant drop of approximately 8% in the advanced stages [294]. Therefore, predicting the response of CRC patients to FOLFOX treatment is closely related to cancer stage. Taking into account this relationship, the prediction task involved combining insights from two trained models: an MLP dedicated to the classification of the FOLFOX response and a CNN designed to classify the stages of cancer. This integration was facilitated through the use of a DL model using bimodalities, as illustrated in Figure 2.7. The bimodal framework allowed for the incorporation of cancer stage information into the prediction model, recognising its crucial role in predicting the response to FOLFOX treatment.

In the context of combining multiple neural networks, two primary methods are early fusion and late fusion. Early fusion involves amalgamating various data modes into a single vector before the model extracts features [295]. On the other hand, late fusion combines models with separate data before merging them into a single model for the final prediction, allowing the use of diverse data modes [296]. In this study, the late-fusion approach was used, where each neural network returned a set of probabilities. These probabilities were combined to derive an adjusted probability, which contributed to the final prediction [296]. The fused model behaved as though it had been applied to a unified dataset, despite being trained on distinct data modes. The output value of the SoftMax activation function was determined as the maximum probability associated with each patient's class, as represented by Equation 2.23.

$$f(x, y) = max(f(x), f(y)) \tag{2.23}$$

Where *f(x,y)* was the overall probability of the subject and its association with a particular class. The maximum probability of the subject and its association with class X and Y.

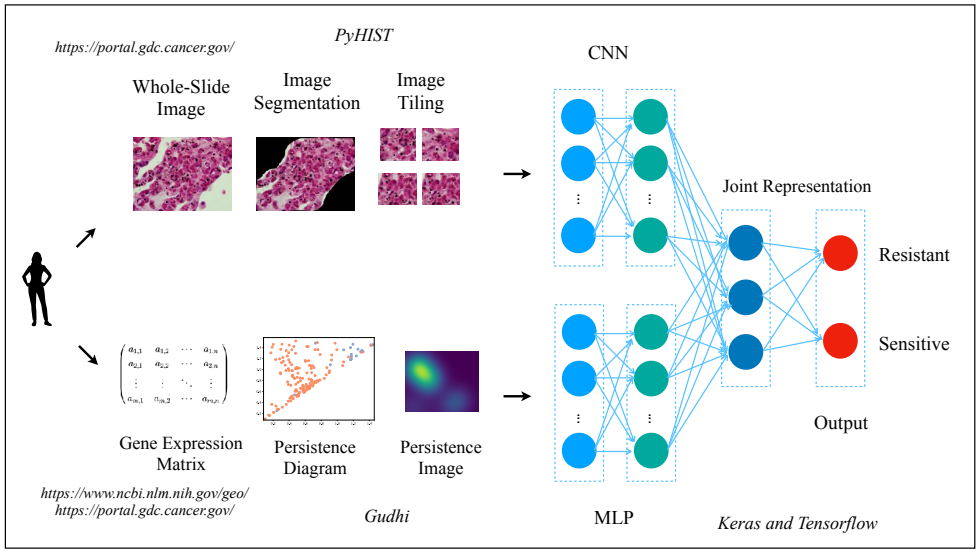


Fig. 2.7 The use of whole slide images (WSIs) and topological summaries for predicting the response to FOLFOX. In this approach, a multilayer perceptron (MLP) was employed for handling topological summaries, while a convolutional neural network (CNN) was used for processing WSI data. The combination of these two modalities enhances the prediction model, enabling a more comprehensive analysis of FOLFOX response. Italicised words indicate computational tools, including *gudhi*, Histological Image Segmentation Tool (*PyHIST*), and *keras and Tensorflow*, used to perform the specified tasks, and data sources for gene expression matrix data, including the National Center for Biotechnology Information-Gene Expression Omnibus (NCBI-GEO) and Genomic Data Commons (GDC), as well as for WSI from GDC.

The input data for the DL model comprised the necessary bimodalities for each patient, involved a single PI and the corresponding WSI with the same cancer stage as the patient. Subsequently, the DL model, constructed to accommodate these bimodalities, was tested, and performance metrics were acquired. The testing of the DL model using bimodalities was performed using the Python programming packages *keras* (version 2.10.0), *Tensorflow* (version 2.9) [280, 281].

Data Organisation for Training and Testing

To prevent data leakage, the following steps were taken. During MLP training, only patients from the combined microarray training subset (that is, the use of PIs) were used. The test subset (that is, the use of PIs constructed from RNA-Seq data) was

kept aside and used to test the DL model using bimodalities. For MobileNet training, only WSI images from the TCGA-COAD/READ cohort were used. Testing the DL in the context of using bimodalities, involved the use of matched WSIs generated from CRC patients with gene expression data generated from the RNA-Seq sample. The number of training and test subsets for CRC patients who are resistant or sensitive to FOLFOX and their corresponding cancer stage is summarised in Table 2.5.

Table 2.5 The number of samples across different cancer stages in the training and testing sets for the *GeTopology* model. The table includes Folinic acid, Fluorouracil and Oxaliplatin (FOLFOX) resistance$_{\text{Train}}$ with 142 stage I and 62 stage IV patients, and FOLFOX sensitive$_{\text{Train}}$ with 72 Stage I and 108 Stage IV patients. For the test set, FOLFOX Resistance$_{\text{Test}}$ has 1 stage I and 7 Stage IV patients, while FOLFOX Sensitive$_{\text{Test}}$ has 6 Stage I and 1 Stage IV patients.

Dataset	**Stage I**	**Stage IV**
FOLFOX Resistance$_{\text{Train}}$	142	62
FOLFOX Sensitive$_{\text{Train}}$	72	108
FOLFOX Resistance$_{\text{Test}}$	1	7
FOLFOX Sensitive$_{\text{Test}}$	6	1

The developed framework was compared to the published framework called *NSCLC*, characterised by the probability of fusion of RNA sequencing and histological imaging, and was selected for structural similarity to the DL model with two modes [199]. *NSCLC's* tools integrated Omics data in the form of RNA-Seq and WSI data for the classification of cancer types. Gene expression data was used to classify patients using a support vector machine (SVM), and CNN classified WSI. The two models were then fused together. In its work published in 2021, *NSCLC* was used to classify adenocarcinoma and squamous cell carcinoma in lung cancers, and the mean macro F1 score was 95% [199]. To test *NSCLC*, the code in the GitHub repository https://github.com/pacocp/NSCLC-multimodal-classification was used and the computational packages used in the framework are summarised in Table 2.6.

Summary

In summary, the methods used involve the classification of the FOLFOX response in CRC patients, using various data types, including microarray and RNA-Seq. Gene expression data were encoded using TDA, where we compared the effectiveness of signed-TOM and distance correlation. Additionally, WSI were processed through image segmentation and tiling to augment the image dataset. TDA, specifically PIs, were

Table 2.6 A comprehensive overview of the bioinformatics tools and data sources used in this study for a variety of bioinformatics tasks. Notable tools and data sources include the National Center for Biotechnology Information-Gene Expression Omnibus (*NCBI-GEO*) for microarray data acquisition, *affy* and *vsn* for microarray data normalisation, and *limma* for quality control and differential gene expression analysis. For RNA-Seq data, the *DESeq2* package was employed for normalisation and differential gene expression analysis. Co-expression analysis and identification of the hub gene were performed using Weighted Gene Correlation Network Analysis (*WGCNA*). Over-representation analysis utilised the *clusterProfiler* package, with reference to Gene Ontology (*GO*) and Kyoto Encyclopedia of Genes and Genomes (*KEGG*). Whole-slide image data (WSI) acquisition was facilitated through the Genomic Data Commons portal (*GDC Data Portal*), while image processing employed the Histological Image Segmentation Tool (*PyHIST*) toolbox. The DL libraries *keras* and *TensorFlow* were used for WSI classification. Additionally, *scikit-learn* was used for persistence image (PI) classification. Topological data analysis (TDA) was performed using the distance correlation package (*dcor*) and *gudhi* packages. ML tasks, including resistance to FOLFOX prediction, were executed using *keras* and *TensorFlow*. Acronyms have been explained for clarity, with italicised resources denoting Bash, R, or Python packages, and others indicating web applications. Each tool is hyperlinked for convenient reference.

Task	**Tools**	**References**
Microarray Data Acquisition	NCBI-GEO	[148]
Microarray Data Normalisation	*affy*, vsn	[85, 297]
Microarray Data QC	*limma*	[83]
Microarray Data Differential Gene Expression	*limma*	[83]
RNA-Seq Data Acquisition	GDC Data Portal	[298]
RNA-Seq Data Normalisation	*DESeq2*	[131]
RNA-Seq Data QC	*DESeq2*	[131]
RNA-Seq Differential Gene Expression	*DESeq2*	[131]
Co-expression Analysis	*WGCNA*	[143, 144]
Hub Gene Identification	*WGCNA*	[143, 144]
Over-representation analysis	*clusterProfiler*, GO, KEGG	[162, 260, 140]
WSI Data Acquisition	GDC Data Portal	[299]
WSI Processing	*PyHIST*	[218]
WSI Classification	*keras*, *TensorFlow*	[280, 281]
TDA	*dcor*, *gudhi*	[275]
PI Classification	*scikit-learn*, *keras*, *TensorFlow*	[291, 280, 281]
Resistance to FOLFOX Prediction	*keras*, *TensorFlow*	[280, 281]

Note: A hyperlink has been added to each tool/data source in the table. Italicised resources denote Bash, R, or Python packages, and other resources denote web applications.

employed to train and test an MLP, while processed WSIs were used to train and test a MobileNet. The integration of these modalities was achieved by combining MLP and MobileNet to enhance the robustness of our predictive model. Rigorous

measures were taken to prevent data leakage during training, and the entire framework was meticulously designed to address the specific nuances of the FOLFOX response classification in patients with CRC. To benchmark the efficacy of our approach, we compared our bimodal framework, named *GeTopology*, with a published approach, *NSCLC*. The following section will present the results and insights gained from the application of these methodologies, shedding light on the nuances of the prediction of the FOLFOX response in patients with CRC. The *GeTopoogy* framework is accessible on GitHub at the following URL: https://github.com/lebomashatola/GeTopology.

Chapter 3

Results

In Chapter 3, the results of the predictive system that classifies FOLFOX resistance, based on a DL model that combines gene expression (Section 2.2.1) and WSI data (Section 2.1.1), are presented. The initial steps involved the identification of hub genes through a combination of differential gene expression analysis and co-expression analysis. This process was undertaken to pre-select the gene expression matrix, followed by the application of TDA to encode the processed gene expression data into a comprehensive topological summary in the form of PIs. Following this, the MLP model underwent training using PIs generated from both signed-TOM and distance correlation measures. Its performance for the classification of CRC patients resistant or sensitive to FOLFOX was analysed in a comparative manner, and the measures were evaluated against each other to identify the optimal preprocessing step.

For the WSIs, preprocessing included image segmentation, and increase in sample numbers was achieved through tiling. The tiled WSIs were then subjected to training using a MobileNet. The integration of MLP and MobileNet models formed the basis for a collective prediction of resistance to FOLFOX in patients with CRC. The effectiveness of the predictive framework for FOLFOX resistance or sensitivity classification was further evaluated through a comparison with a similar method denoted *NSCLC*. The robustness and generalisability of the framework was further examined by applying it to additional cancer datasets. The subsequent sections delve into the specific results and findings derived from these analyses.

3.1 Differential Gene Expression Analysis

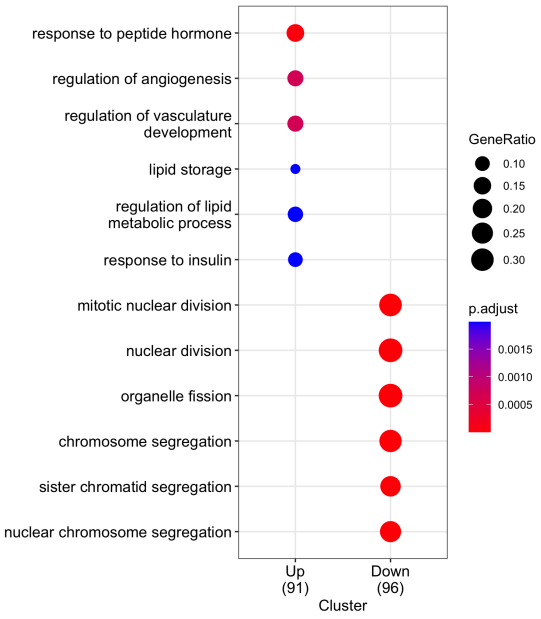
In exploring the distinctions in gene expression and read counts between patients resistant to FOLFOX and those exhibiting sensitivity, the identification of genes that achieve statistical significance is crucial. This analysis serves as a crucial step in identifying genes that underscore the biological variances between the two states of the disease. The application of *DESeq2* to RNA-Seq data and *Limma* to microarray data facilitated this comparative analysis. The following section delineates the results of this analysis, illustrating the number of genes identified as up-regulated and down-regulated for each dataset. These DEGs subsequently formed the basis for the pre-selection of the gene expression matrix in each dataset, paving the way for comprehensive co-expression analysis. Table 3.1 provides information on the differential gene expression patterns observed in the specified datasets.

Table 3.1 The datasets, namely GSE69657, GSE81653, GSE72968, and RNA-Seq, are characterised by the counts of up-regulated and down-regulated genes. For GSE69657, there are 123 genes up-regulated and 115 genes down-regulated. In the GSE81653 dataset, 230 genes exhibit up-regulation, while 210 genes are down-regulated. Similarly, GSE72968 shows 119 up-regulated genes and 113 down-regulated genes. The RNA-Seq dataset, displays 150 up-regulated genes and 210 down-regulated genes. A relatively similar number of DEGs are observed across analysed datasets. This table serves as a concise overview of the differential gene expression patterns observed in the specified datasets, shedding light on the regulatory dynamics of genes within each context.

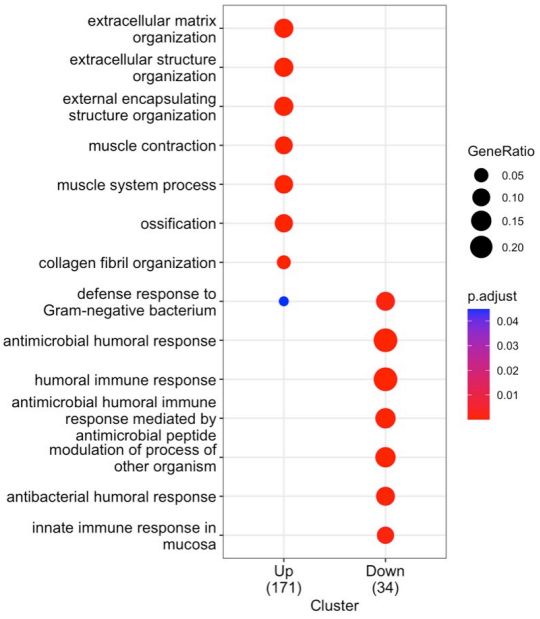
Dataset	**Up-regulated Genes**	**Down-regulated Genes**
GSE69657	123	115
GSE81653	230	210
GSE72968	119	113
RNA-Seq	150	210

The investigation into resistance to FOLFOX involved a critical step of over-representation analysis, which aimed to unveil over-represented biological annotations within the DEGs. Initial identification of DEGs involved a comparison of gene expression profiles between conditions, specifically, the FOLFOX resistant and FOLFOX sensitive states. To distinguish enriched biological annotations among the DEGs, over-representation analysis was used. This analytical strategy utilised databases including GO and pathway databases including KEGG to assign functional annotations to genes. In particular, the analysis applied a logFC threshold of 0.5 for the identification of genes regulated by the up (logFC $>$ 0.5) and down (logFC $<$ -0.5), coupled with a cut-off 5% for the BH adjusted p-value. The results of this analysis, which detail the

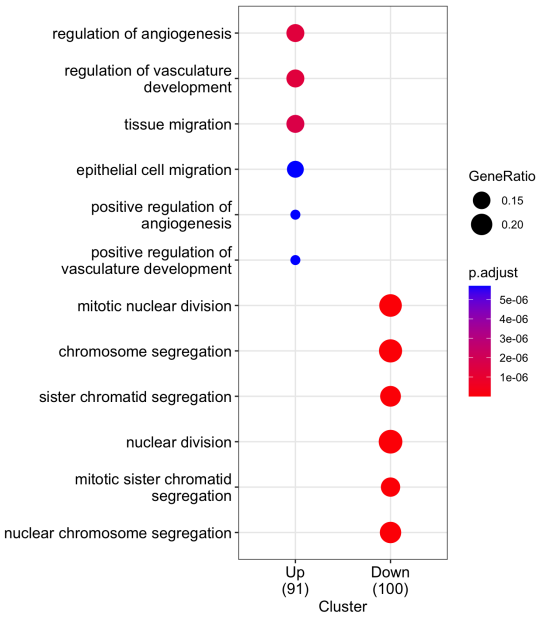
enriched biological annotations, played a pivotal role in deciphering the molecular underpinnings of resistance to FOLFOX (Figure 3.1).



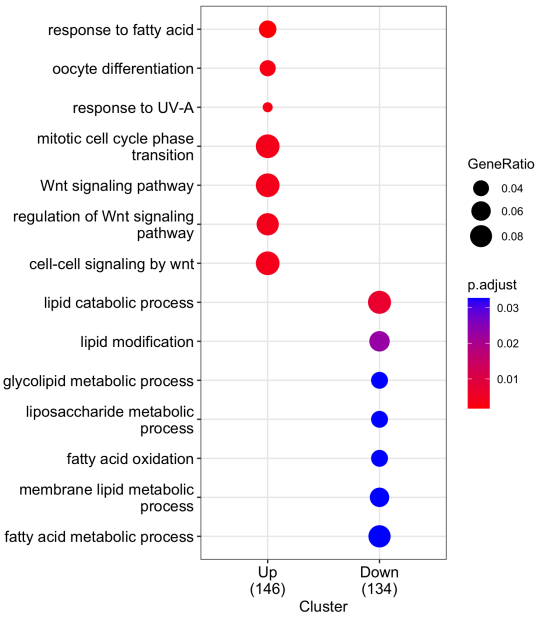
GSE69657



GSE81653



GSE72968



RNA-Seq

Fig. 3.1 The figure depicts the results of over-representation analysis on DEGs. The x-axis represents the number of genes over-represented in GO for each gene module, while the y-axis illustrates the enrichment of GO terms. Each point on the graph corresponds to a specific GO term associated with a gene module. The colour of the points indicates the BH-adjusted p-value, with lower values indicating higher statistical significance. Additionally, the size of the points reflects the gene ratio, representing the percentage of total DEGs within the given GO term. Larger points indicate a higher gene ratio, highlighting the proportion of genes from the input dataset that map to the enriched GO term. This comprehensive visualisation provides insights into both the statistical significance and the biological relevance of DEGs enrichment in the analysed dataset. Interestingly, overlapping GO terms between the analysed datasets include upregulation of angiogenesis and vasculature development, as well as downregulation of mitotic nuclear division and lipid and fatty acid metabolic pathways. The GSE81653 dataset exhibited distinct results, particularly in the downregulation of immune system-related GO terms and upregulation of extracellular structure organisation.

3.2 Co-expression Analysis

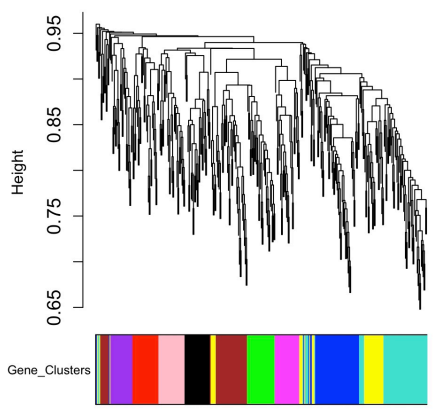
Co-expression analysis is a widely used technique aimed at revealing regulatory networks and elucidating transcription profiles through clustering. This approach is crucial for identifying hub genes and understanding their biological functions. The determination of the optimal power is crucial and the scale-free topology is assessed for values ranging from 1 to 20. In the datasets employed, which were pre-processed by pre-selecting the identified DEGs, the power values that best correspond to the scale-free topology were carefully selected, considering a range between 5 and 8. These optimal power values, along with their respective evaluations, are presented in Table 3.2. This systematic approach ensures that the co-expression analysis is conducted under conditions that promote a scale-free network structure, enhancing the reliability of the derived regulatory networks and facilitating a more accurate understanding of hub genes and their associated biological functions.

Following the calculation of gene pair correlations using the signed-TOM approach within the *WGCNA* framework, the resulting correlations were organised to identify distinct gene modules. A gene dendrogram, depicted in Figure 3.2, visually represents and evaluates the hierarchical relationships among these modules. The generation of the gene dendrogram involved essential steps within *WGCNA*, including the calculation of the adjacency matrix, highlighting network connections between genes. Subsequently, the matrix was transformed into a signed TOM, accentuating the interconnectedness

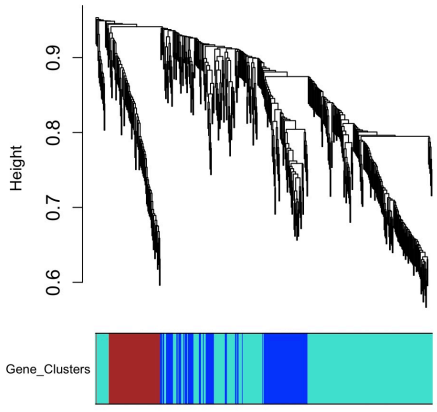
Table 3.2 The power values, representing the soft-thresholds, were carefully selected for different datasets (GSE69657, GSE81653, GSE72968, and RNA-Seq). The assessment includes a scale-free topology, indicating the extent to which the network conforms to a scale-free structure, and mean connectivity, representing the average number of connections per gene in the network. Insights into the optimal soft-thresholding conditions for co-expression analysis, crucial for constructing robust regulatory networks and understanding the biological functions of hub genes are presented. Noteworthy observations include GSE72968 exhibiting a high scale-free topology (0.88) and mean connectivity (104). GSE81653 showing the lowest scale-free topology (0.77) and mean connectivity (23).

Dataset	**Power**	**Scale-Free Topology**	**Mean Connectivity**
GSE69657	6	0.80	27
GSE81653	8	0.77	23
GSE72968	6	0.88	104
RNA-Seq	5	0.85	102

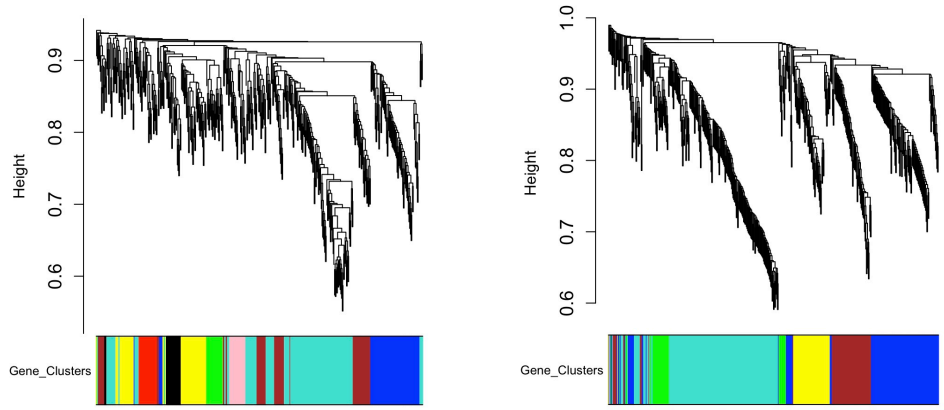
of genes within the modules. The hierarchical clustering of genes based on this TOM led to the construction of the gene dendrogram. This dendrogram provides a graphical illustration of the relationships among genes within modules, facilitating the interpretation of their hierarchical organisation based on correlated expression patterns. These steps are pivotal for unraveling the modular structure of co-expressed genes and gaining insights into the intricate relationships within a biological system.



GSE69657



GSE81653

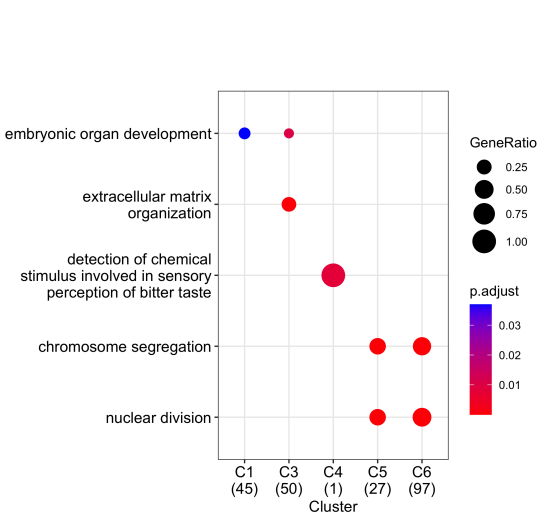


GSE72968 RNA-Seq

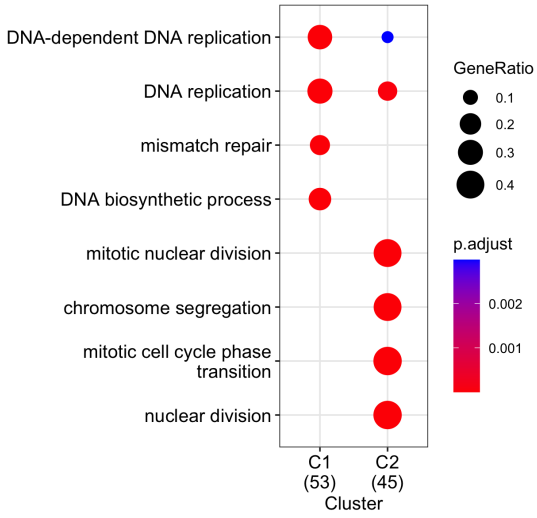
Fig. 3.2 The gene dendrogram showing the hierarchical tree delineated by coloured blocks, each corresponding to a distinct gene module. This clustering was generated using the *WGCNA* framework, incorporating the signed-TOM approach to evaluate the correlations of the gene pairs. The hierarchical organisation of modules within the gene dendrogram provides valuable insight into the relationships among co-expressed genes, contributing to a deeper understanding of the modular structure and intricate relationships within the biological system. Numerous gene modules were shown, with GSE81653 showing the least number of gene modules (3) and GSE72968 showing the most number of gene modules (15).

The gene selection process involved choosing the top 15% of genes with the highest estimated k_{ME} (eigengene-based connectivity). This heuristically selected subset of genes was utilised to pre-select the gene expression matrix. Table 3.3 provides a summary of the mean k_{ME} values for hub genes identified in each dataset. This table offers a concise overview of connectivity measures based on eigengenes associated with the hub genes in the datasets considered.

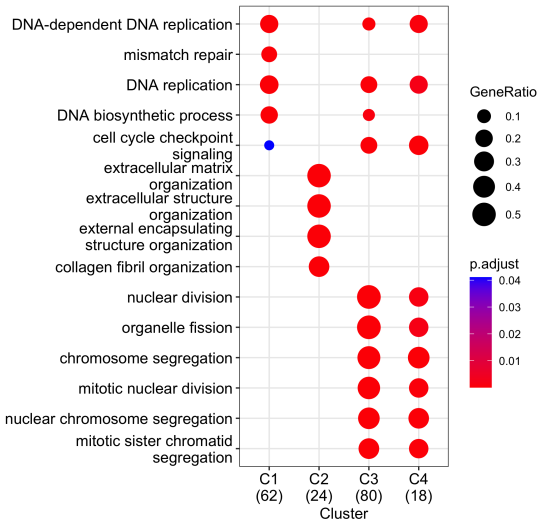
An exploration of the biological functions associated with the hub genes was carried out by analysing their over-representation, as illustrated in Figure 3.3. This examination involved evaluating the terms of the ontology of the enriched gene, providing insight into the functional roles and processes in which the identified hub genes are involved. Significantly enriched GO terms were selected based on a BH-adjusted p-value threshold of less than 5%, ensuring the identification of robust and statistically significant functional annotations associated with the hub genes. This stringent criterion improves the reliability of over-representation analysis, shedding light on the key biological processes in which the hub genes play a pivotal role.



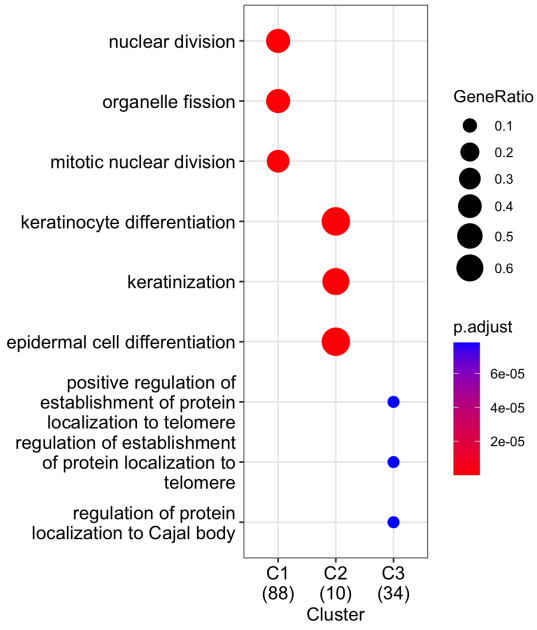
GSE69657



GSE81653



GSE72968



RNA-Seq

Fig. 3.3 The *x-axis* represents the quantity of hub genes within each module, labeled as "C," and the gene module number, while the *y-axis* illustrates the enriched Gene Ontology (GO) terms. Point color indicates the Benjamini-Hochberg (BH)-adjusted p-value, with a color gradient representing the significance level. Point size corresponds to the gene ratio, defined as the proportion of total hub genes associated with the respective GO term. The gene ratio is calculated as the number of mapped genes to the specific GO term from the input dataset. Larger points indicate a higher gene ratio. The selection of significantly enriched GO terms is based on a stringent criterion of BH-adjusted p-value less than 5%, ensuring robust and statistically significant functional annotations associated with the hub genes. Gene modules showed overlapping enrichment for DNA repair mechanisms, DNA replication, nuclear division and mitosis related biological processes. These shared biological processes suggest a potential core set of genes that play crucial roles in fundamental aspects of cell regulation and division.

Table 3.3 The mean k_{ME} values for hub genes identified in different datasets including the GSE69657, GSE81653 and GSE72968 from NCBI-GEO and RNA-Seq from GDC. The top 15% genes with the highest estimated eigengene-based connectivity (k_{ME}) were selected during the gene selection process. The table includes information on the mean k_{ME}, the number of hub genes, and the number of gene modules for each dataset. The k_{ME} measures associated with hub genes, providing insights into their central role within co-expression networks in the considered datasets are summarised. Notably, GSE72968 stands out with a mean k_{ME} of 0.77, indicating the large extent of centrality/connectivity of hub genes in this dataset. GSE81653 exhibits a lower connectivity (0.57) of hub genes compared to all other datasets in the analysis.

Dataset	**Mean-k_{ME}**	**Number of Hub Genes**	**Number of Gene Modules**
GSE69657	0.71	108	19
GSE81653	0.57	56	25
GSE72968	0.77	138	18
RNA-Seq	0.73	124	18

Through an examination of the biological functions inherent in the datasets, shared cancer-related GO terms were identified and summarised in Table 3.4. In Figure 3.3, gene modules that demonstrate similar biological functionalities were labelledd on the y-axis. This integration of functional information provides a comprehensive view of commonalities in biological processes across the analysed datasets and their corresponding gene modules. The C1-5 gene modules, selected from each dataset, were chosen due to overlap of biological functions. The significance of identifying overlapping biological processes, represented by the gene modules C1-C5 across different datasets, lies in the robustness and consistency of these observations. When similar biological functions are independently observed under different experimental conditions or datasets, it strengthens confidence in the validity and reproducibility of those findings. This convergence across multiple experiments enhances our understanding of shared regulatory mechanisms or pathways underlying the studied biological processes, contributing to a more comprehensive and reliable interpretation of the data.

3.3 Topological Summaries

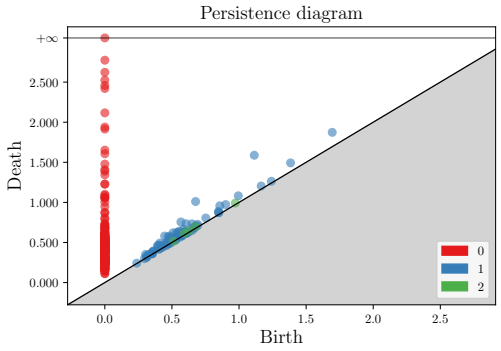
The gene expression matrix underwent pre-selection using hub genes, determined through a combination of differential gene expression and co-expression analysis, followed by encoding into a topological signature (PI) using TDA for each patient. The input data were initially transformed into a correlation matrix, employing techniques

Table 3.4 For each dataset (GSE69657, GSE81653, GSE72968, and RNA-Seq), the table details the gene modules (labelled as C1-C4 in Figure 3.3) and their associated enriched biological functions. These modules are associated with specific enriched biological functions, providing insights into the shared regulatory landscape across datasets. For example, GSE69657 exhibits overlaps in nuclear division functions in modules C5 and C6, while GSE81653 shows enrichment in mismatch repair and nuclear division in modules C1 and C2. In GSE72968, modules C1, C3, and C4 display enrichments in mismatch repair, cell cycle checkpoint, and nuclear division. RNA-Seq data reveals enrichment in nuclear division in module C1. Focussing on these shared biological processes is crucial as it helps to uncover consistent patterns in different experiments.

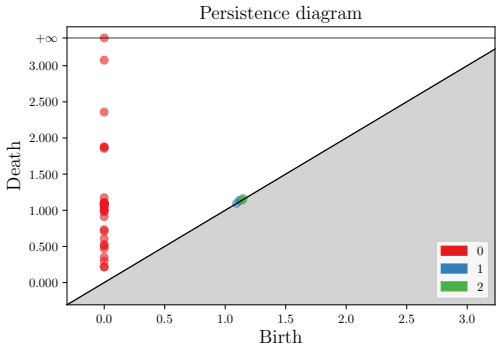
Dataset	**Gene Module**	**Enriched Biological Function**
GSE69657	C5	Nuclear division
	C6	Nuclear division
GSE81653	C1	Mismatch repair
	C2	Nuclear division
GSE72968	C1	Mismatch repair/Cell cycle checkpoint
	C3	Mismatch repair/Nuclear division
	C4	Mismatch repair/Nuclear division
RNA-Seq	C1	Nuclear division

Note: C1-C4 are related to the gene modules labelled on the x-axis in Figure 3.3.

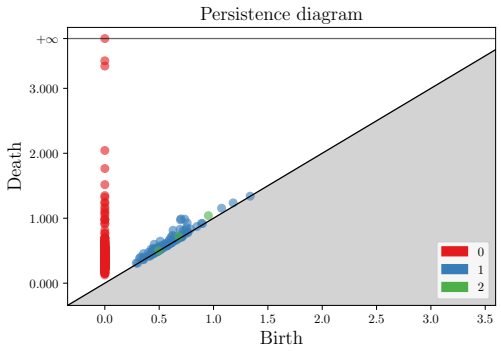
such as distance correlation. Additionally, a signed-TOM method was used, and a comparison was conducted to assess which input transformation technique resulted in a more informative topological signature. The topological signature, which consists of a combination of *Betti* numbers, is visually represented by a PD, as depicted in Figure 3.4. This PD encapsulates patient-specific topological features derived from the gene expression matrix, offering a unique perspective on the underlying biological structure. During this process, a comparison of the PDs constructed from signed-TOM and distance correlation measures was conducted. The objective was to choose the method that resulted in numerous and diverse topological features, as indicated by more data points on the PD representing *Betti*-0, *Betti*-1 and *Betti*-2 homology groups. This comparison allowed for the selection of the most suitable input transformation technique, which ultimately enhanced the discriminative power and interpretability of the topological signature, facilitating deeper insights into the biological structure underlying the gene expression data.



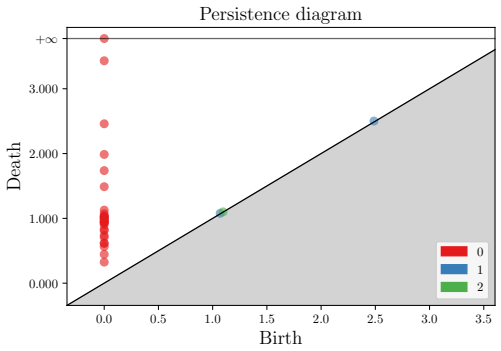
GSE69657 - Signed-TOM



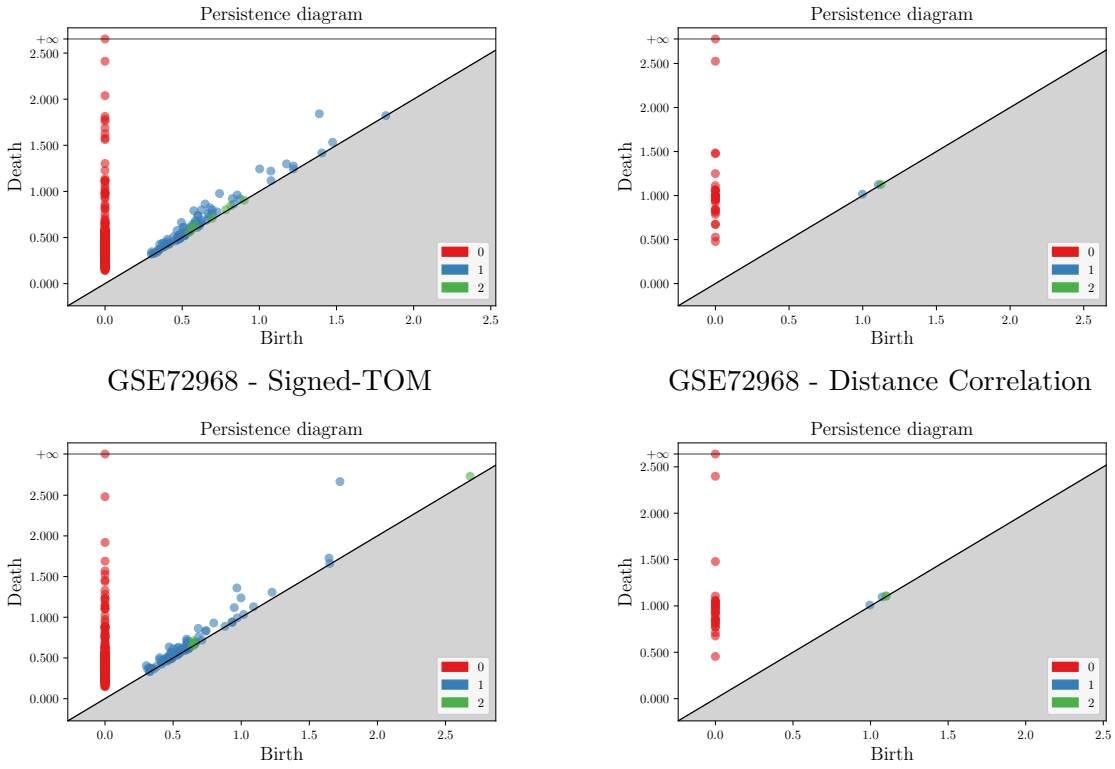
GSE69657 - Distance Correlation



GSE81653 - Signed-TOM



GSE81653 - Distance Correlation



GSE72968 - Signed-TOM

GSE72968 - Distance Correlation

RNA-Seq - Signed-TOM

RNA-Seq - Distance Correlation

Fig. 3.4 A representative PD from a patient created using signed-TOM measurements and distance correlation measures. The x-axis represents the birth parameter (b), and the y-axis represents the death parameter (d). Topological features, classified into homology groups, are depicted as points: *Betti*-0 (in red), *Betti*-1 (in blue), and *Betti*-2 (in green). The grey area denotes regions where no topological features were identified (i.e., where b is less than d). Notably, the PDs on the left (constructed from Signed-TOM) exhibits a greater number of *Betti* numbers, indicative of a more intricate and diverse topological signature. In contrast, the PDs on the right shows (constructed distance correlation measures) a simpler topological structure with fewer *Betti* numbers. This simplicity implies a less complex and more straightforward topological configuration in the data represented on the right.

3.4 Predictive Learning

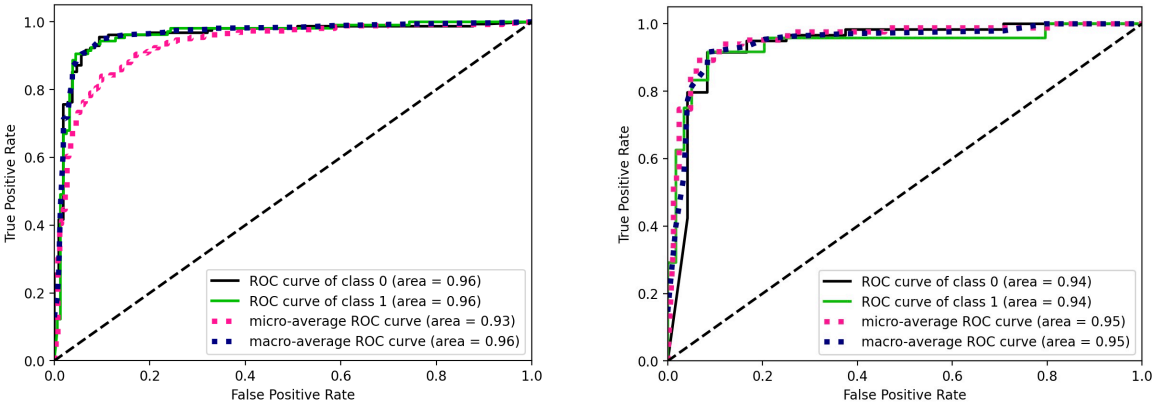
3.4.1 Image Classification

The prediction of cancer stage, distinguishing between stage one and stage four, was performed using processed WSI data from TCGA-COAD/READ patients. Before model training, WSI preprocessing involved image segmentation and tiling. The images were segmented to extract relevant regions and subsequent tiling was performed to divide the WSI into smaller tiles, facilitating more efficient model training. A CNN model, specifically using the MobileNet architecture, was employed for training the tiled images. The performance of the model was evaluated using various metrics, including TPR, Precision, macro F1-scores, and overall Accuracy. These model metrics are concisely presented in Table 3.5. In particular, the model was trained on 70% of the WSI tiles and tested on the remaining 30% to ensure robust evaluation and validation of its predictive capabilities. It is crucial to note that the WSI used in this case were from the TCGA-COAD/READ cohort, encompassing both patients with cancer and healthy individuals.

Table 3.5 Performance metrics of the cancer stage prediction model evaluated on whole slide images (WSIs) data TCGA-COAD/READ patients. The model was trained on 70% of WSI tiles (WSI_{Train}) and tested on the remaining 30% (WSI_{Test}). Performance metrics include true positive rate (TPR), precision, F1-score (Macro), and overall accuracy, each expressed with mean values and standard deviations (±) for robust evaluation. Information on the effectiveness of the model to distinguish between cancer stages is presented, with performance metrics calculated for both the training and the testing data sets. The TPR reflects the percentage of correctly identified positive instances, precision indicates the accuracy of positive predictions, F1-score (Macro) considers both precision and recall for a balanced evaluation, and accuracy provides an overall measure of correct predictions. For WSI_{Train}, the model demonstrates a TPR of 93.32% ± 0.78, precision of 94.83% ± 0.83, F1-score (Macro) of 94.07% ± 0.98, and an accuracy of 94.58% ± 0.87. Similarly, for WSI_{Test}, the model shows a TPR of 90.11% ± 1.41, precision of 91.91% ± 1.24, F1-score (Macro) of 91.01% ± 1.23, and an accuracy of 91.51% ± 1.62. These metrics collectively offer a comprehensive assessment of the model's performance in cancer stage prediction on WSIs.

Method	**TPR**	**Precision**	**F1-Score (Macro)**	**Accuracy**
WSI_{Train}	93.32 ± 0.78	94.83 ± 0.83	94.07 ± 0.98	94.58 ± 0.87
WSI_{Test}	90.11 ± 1.41	91.91 ± 1.24	91.01 ± 1.23	91.51 ± 1.62

ROC curves evaluate MobileNet efficacy in predicting cancer stages, particularly in the context of WSI from the TCGA-COAD/READ cohort. These curves visually depict the model's discrimination ability, illustrating sensitivity and specificity across diverse classification thresholds. Crucially, ROC curves enable a direct comparison between MobileNet's training and testing performances, serving as a tool to assess generalisation to unseen data and identify potential overfitting. The analysis of the AUC concisely summarises the overall performance of the MobileNet for the TCGA-COAD/READ cohort (Figure 3.5).



WSI - Train WSI - Test

Fig. 3.5 The receiver operating characteristic (ROC) curves illustrate the performance of a MobileNet in the classification of cancer stages, with the FPR plotted on the x-axis and the TPR on the y-axis. The curves illustraate the model's ability to balance sensitivity and specificity across different threshold values. F1-scores, both micro and macro, along with the AUC, are reported for each predicted class-cancer stage one and cancer stage four. The F1-scores provide insights into the model's precision and recall, while the AUC quantifies the overall discriminative power in classifying different cancer stages. The AUC values, are reported as 0.96 for the training set and 0.95 for the testing set. The AUC values near 1 indicate high predictive accuracy, with the training and testing sets demonstrating robust discriminative power.

3.4.2 Topological Summaries Classification

An integral step in TDA involves inputting data in the form of correlation measures to construct the VR complex. Typically, distance correlation measures are employed in TDA, capturing both linear and non-linear relationships between variables. Initially, TDA using distance correlation measures to determine PI had limited predictive power. This led to the exploration of the use of signed-TOMs as input to TDA. Signed-TOM

takes into account neighbourhood connections by providing information on direct and indirect connections [143]. PIs generated for patients with resistant or sensitive CRC to FOLFOX were classified using an MLP. The model performance was evaluated using TPR, Precision, F1-score (Macro) and accuracy. The MLP was trained in the samples described in Table 2.4 (referred to as *MLP training*). Subsequently, the MLP was tested on datasets summarised in Table 2.3. Prediction performance using an MLP was evaluated, and performance metrics are summarised in Table 3.6.

Table 3.6 Performance metrics for FOLFOX response prediction using distance correlation (DCOR) and signed-topological overlap matrix (Signed-TOM). The evaluated methods include DCOR MLP and signed-TOM MLP for both training and testing datasets. True positive rate (TPR), precision, F1-score (Macro), and accuracy are reported, with results presented as a mean with the standard deviation (±). In the training phase, the DCOR MLP demonstrates an accuracy of 80.02% ± 1.46. However, in the testing phase, the DCOR MLP exhibits a lower accuracy of 61.21% ± 0.92. These contrasting results suggest variations in model performance between training and testing datasets, hinting at potential overfitting effects. In contrast, the Signed-TOM MLP demonstrates improved performance in testing with an accuracy of 78.05% ± 0.79. These findings highlight the signficance of data pre-processing methods (i.e. signed-TOM and distance correlation measures) affecting DL classification tasks.

Method	**TPR**	**Precision**	**F1-Score (Macro)**	**Accuracy**
DCOR MLP_{train}	81.74 ± 1.21	80.13 ± 1.62	79.81 ± 1.45	80.02 ± 1.46
Signed-TOM MLP_{train}	81.24 ± 1.45	82.00 ± 1.24	81.36 ± 1.34	80.96 ± 1.11
DCOR MLP_{test}	63.11 ± 1.21	67.89 ± 1.56	65.50 ± 1.22	61.21 ± 0.92
Signed-TOM MLP_{test}	79.42 ± 1.11	78.11 ± 0.93	78.77 ± 0.86	78.05 ± 0.79

ROC curves are provided to illustrate the performance of the MLP model in classifying patients with CRC as resistant or sensitive to FOLFOX treatment. The input topological summaries were generated using either distance correlation measures or signed-TOM. Figure 3.6 visually represents these ROC curves, offering valuable insight into the model's capacity to discern between patients exhibiting varied responses to FOLFOX treatment.

3.4.3 Classification using Deep Learning Model Built on Bimodalities

The *GeTopology* framework was developed employing a pre-trained CNN based on MobileNet. This MobileNet was trained using WSI data from the NCI. Additionally, a

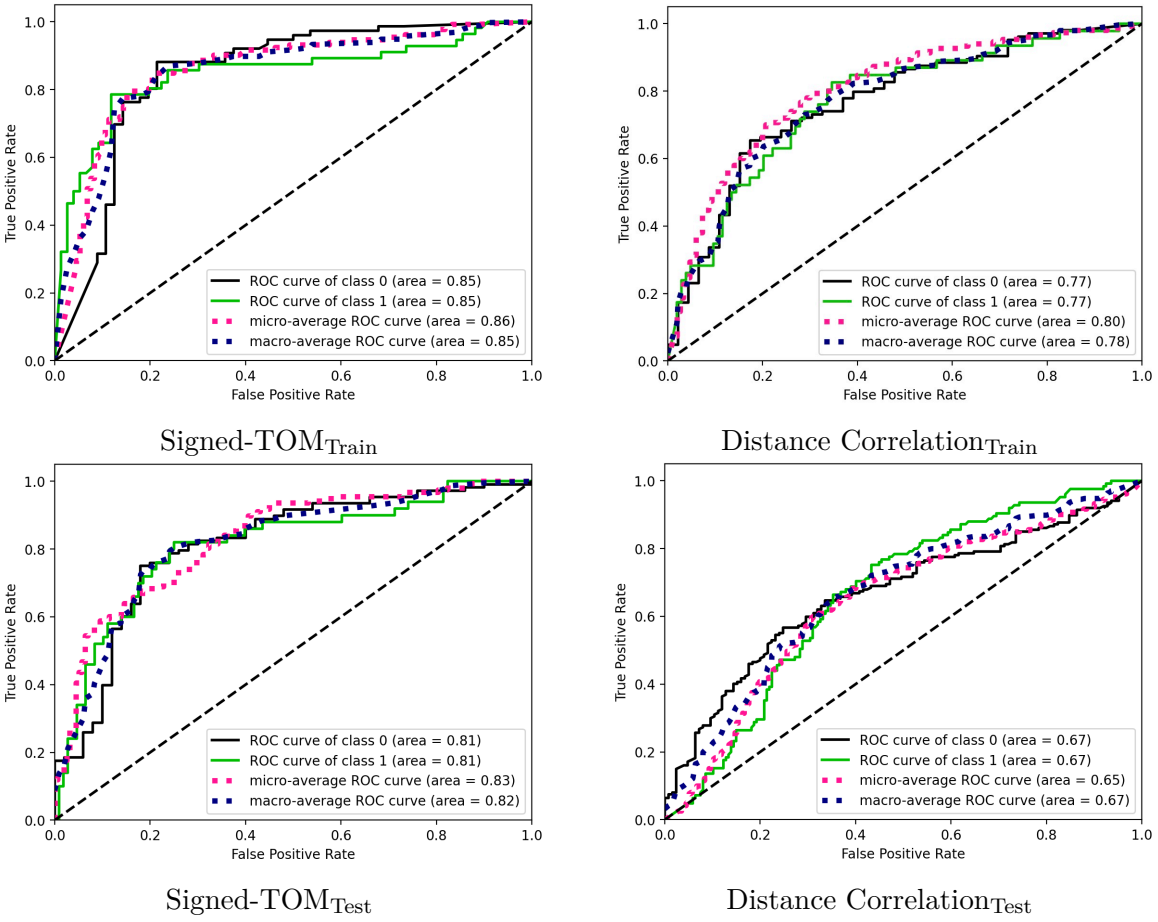


Fig. 3.6 The ROC curves compare signed-TOM and distance correlation measures in distinguishing between FOLFOX-resistant and FOLFOX-sensitive CRC patients. The x-axis represents the FPR, while the y-axis represents the TPR. F1-scores, both micro and macro, are reported together with the AUC for each predicted class. These metrics provide insight into the precision, recall, and overall accuracy of the model to identify resistance to FOLFOX in patients with CRC. The AUC values for the training set are 0.85 for Signed-TOM and 0.78 for distance correlation. Similarly, for the testing set, the AUC values are 0.82 for Signed-TOM and 0.67 for distance correlation. The higher AUC values for Signed-TOM suggest its superior performance in both training and testing scenarios compared to distance correlation. These differences emphasise the importance of selecting an effective data pre-processing method, with Signed-TOM demonstrating enhanced accuracy and robustness in FOLFOX response classification.

pre-trained MLP was used, trained on PI scores generated from merged microarray data. Evaluation of the *GeTopology* framework involved the use of RNA-Seq data from the NCI, together with its corresponding matched WSI. This comprehensive approach integrates both image-based and molecular data to assess the performance and effectiveness of the *GeTopology* framework in the capture and characterisation of complex biological information. Performance metrics such as TPR, precision, F1-score (Macro), and accuracy were used to evaluate the effectiveness of the model shown in Table 3.7.

Table 3.7 Performance metrics for the *GeTopology* framework, evaluating its effectiveness in classifying patients with CRC as resistant or sensitive to FOLFOX treatment. The true positive rate (TPR), precision, F1-Score (macro), and accuracy are reported for the training and testing phases. The models evaluated include topological data analysis used in a multilayer perceptron layer (TDA-MLP), the convolutional neural network (CNN) used to classify WSI, and the proposed *GeTopology*. Results are provided as a mean with the standard deviation (±) for both the training (Train) and testing (Test) datasets, including TPR, precision, F1-Score (Macro) and precision. The table provides a comprehensive overview of the performance metrics for the evaluated models. In the training phase, TDA-MLP shows an accuracy of 80.96% ± 1.11. MobileNet demonstrates superior performance with an accuracy of 94.58% ± 0.87. The proposed *GeTopology* exhibits competitive results with an accuracy of 84.32% ± 1.69 during training. In the testing phase, TDA-MLP demonstrates an accuracy of 78.05% ± 0.79. MobileNet maintains strong performance with an accuracy of 91.51% ± 1.62. The proposed *GeTopology* framework continues to exhibit competitive metrics with an accuracy of 83.65% ± 0.95. These results offer a detailed comparison of how well the models perform in classifying CRC patients by FOLFOX response. Importantly, MobileNet performs a cancer stage task whilst TDA-MLP and *GeTopology* perform CRC patient classification for FOLFOX resistant/sensitive patients, this shows that the high performance of MobileNet enhances *GeTopollogy*s. Moreover, the findings highlight that *GeTopology* is a promising and competitive approach for the task of FOLFOX response classification.

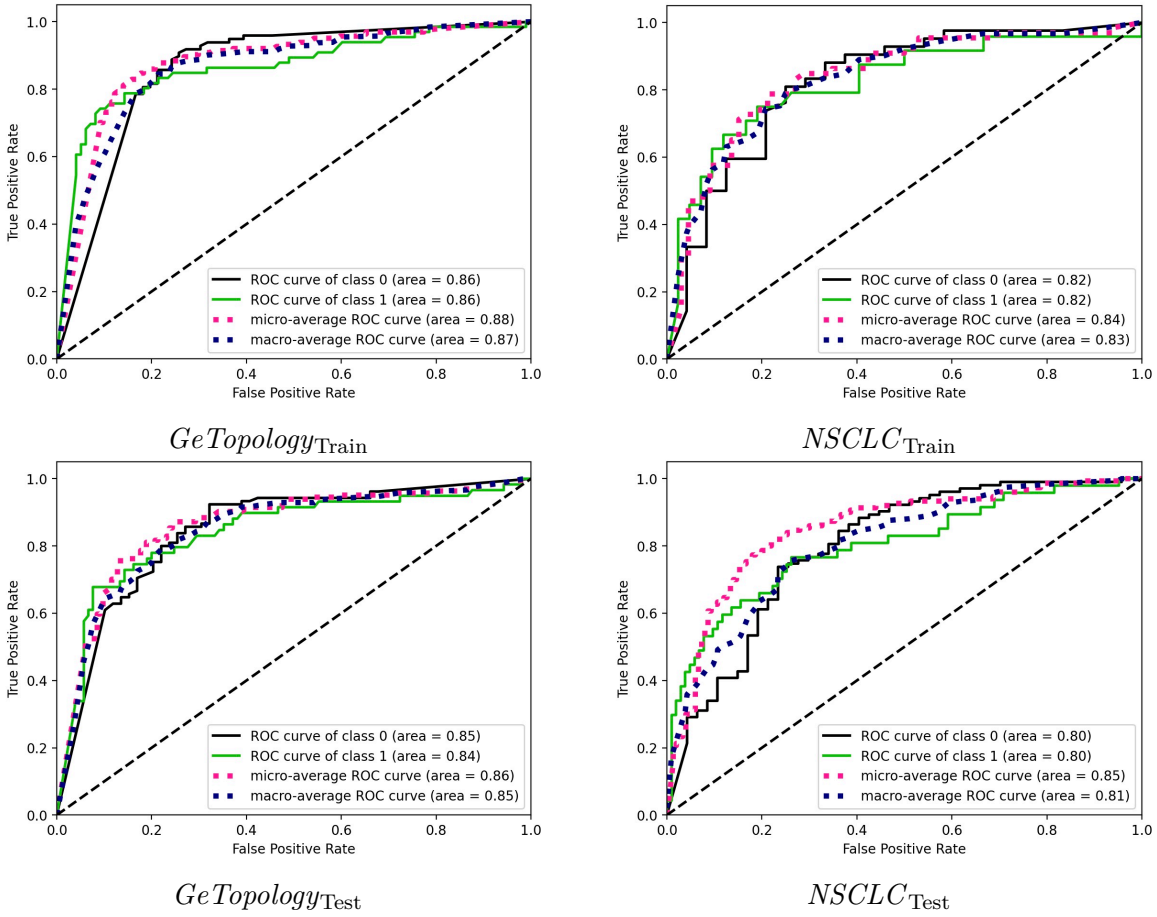
Method	**TPR**	**Precision**	**F1-Score (Macro)**	**Accuracy**
$\text{TDA-MLP}_{\text{Train}}$	81.24 ± 1.45	82.00 ± 1.24	81.36 ± 1.34	80.96 ± 1.11
$\text{MobileNet}_{\text{Train}}$	93.32 ± 0.78	94.83 ± 0.83	93.25 ± 0.98	94.58 ± 0.87
$\textit{GeTopology}_{\text{Train}}$	84.92 ± 1.52	83.91 ± 1.45	84.49 ± 1.15	84.32 ± 1.69
$\text{TDA-MLP}_{\text{Test}}$	79.42 ± 1.11	78.11 ± 0.93	78.76 ± 0.86	78.05 ± 0.79
$\text{MobileNet}_{\text{Test}}$	90.11 ± 1.41	91.91 ± 1.24	91.01 ± 1.23	91.51 ± 1.62
$\textit{GeTopology}_{\text{Test}}$	83.72 ± 1.09	82.98 ± 1.02	83.35 ± 1.00	83.65 ± 0.95

The performance of the *GeTopology* framework was systematically compared to that of *NSCLC*. A detailed evaluation of both models was performed, using specific metrics such as TPR, precision, and macro F1-scores. The summarised results of this comparative assessment are presented in Table 3.8. This table offers a comprehensive overview of performance metrics, allowing an informed comparison of the *GeTopology* framework with the *NSCLC* model. The metrics provide insight into the effectiveness of the models in terms of sensitivity, precision, and overall predictive accuracy.

Table 3.8 Performance metrics of the *GeTopology* framework with those of *NSCLC* (*Non-Small Cell Lung Cancer Phenotype Prediction*) for the classification of CRC patients resistant or sensitive to FOLFOX. Abbreviations: TPR - True Positive Rate. Results are presented as a mean with the standard deviation (±) for both the training (Train) and testing (Test) datasets, including TPR, Precision, F1-Score (Macro), and Accuracy. The *GeTopology* model is evaluated along with the *NSCLC* model, providing information on its relative efficacy in predicting cancer stage. In the training phase, *GeTopology* shows an accuracy of 86.78% ± 1.12. In comparison, *NSCLC* demonstrates an accuracy of 88.01% ± 1.46 during training. In the testing phase, *GeTopology* maintains robust performance with an accuracy of 83.65% ± 0.95. Conversely, *NSCLC* exhibits an accuracy of 81.90% ± 0.84. These results offer a comparative perspective on the performance of *GeTopology* and *NSCLC* in the context of CRC patient classification for FOLFOX response, highlighting the effectiveness of *GeTopology* in classifying CRC patients based on their FOLFOX response.

Method	**TPR**	**Precision**	**F1-Score (Macro)**	**Accuracy**
$GeTopology_{\text{train}}$	87.89 ± 1.11	86.13 ± 1.56	85.67 ± 0.98	86.78 ± 1.12
$NSCLC_{\text{train}}$	88.45 ± 1.93	87.78 ± 1.35	88.21 ± 1.12	88.01 ± 1.46
$GeTopology_{\text{test}}$	83.72 ± 1.09	82.98 ± 1.02	83.35 ± 1.00	83.65 ± 0.95
$NSCLC_{\text{Test}}$	81.32 ± 0.85	80.98 ± 0.78	81.15 ± 0.83	81.90 ± 0.84

The classification performance of the *GeTopology* and *NSCLC* models for CRC patients' sensitivity or resistance to FOLFOX treatment are visually represented through ROC curves in Figure 3.7. These curves offer a comprehensive assessment of the discriminatory ability of models in different response categories. Consequently, ROC and AUC are crucial for a detailed evaluation of the predictive efficacy of both the *GeTopology* and *NSCLC* models in the context of predicting the response to FOLFOX treatment for patients with CRC.



$GeTopology_{\text{Train}}$ $NSCLC_{\text{Train}}$

$GeTopology_{\text{Test}}$ $NSCLC_{\text{Test}}$

Fig. 3.7 The ROC curves compare the performance of *GeTopology* and *NSCLC* in classifying CRC patients resistant to FOLFOX and those sensitive to FOLFOX. The x-axis represents the FPR and the y-axis represents the TPR. F1-scores, both micro and macro, along with the AUC, are provided for each predicted class. These metrics provide insight into the precision, recall and overall discriminative power of *GeTopology* and *NSCLC* in identifying FOLFOX resistance in CRC patients. The AUC values for the training set are 0.87 for *GeTopology* and 0.83 for *NSCLC*. Similarly, for the testing set, the AUC values are 0.85 for *GeTopology* and 0.81 for *NSCLC*. The ROC curves visually depict the trade-off between TPR and FPR, emphasising the discriminatory ability of each model. Higher AUC values indicate better discriminative power, with *GeTopology* demonstrating superior performance over *NSCLC* in both training and testing scenarios.

Summary

In summary, our investigation of the performance of the novel patient classification framework, which incorporates data from different cohorts of patients, has yielded promising results. The ROC curves in Figures 3.7 demonstrate the robust discriminatory capacity of the model, particularly evident in the testing datasets. The key contributions of this research lie in the successful development of a DL model with high TPR and precision in cancer classification. In the upcoming interpretation and discussion sections, the significance of these findings is discussed, study limitations are addressed, and potential applications are explored. In addition, the broader impact of our research on cancer classi and outline avenues for future investigation. In general, these results establish a foundation for understanding the performance of the patient classification framework and its implications for advancing cancer diagnostics.

Chapter 4

Discussion and Conclusions

This study introduces a novel methodology that integrates PH for predictive modelling in CRC, specifically in the context of FOLFOX resistance. The approach identifies the hub genes that govern DNA repair and nuclear division, providing molecular insights for targeted therapies. Combining genetic expression data from PIs with morphological information from WSIs improves predictive accuracy, illustrating clinical relevance in the CRC stage and FOLFOX response predictions. Despite computational challenges, the study justifies its approach through superior predictive performance. Recommendations include biological validation and user-friendly interfaces for broader applicability, significantly advancing precision medicine in CRC. The present study aims to establish a novel framework for predicting drug resistance in CRC patients through a multifaceted approach that incorporates several key components. The methodology presented involves the use of gene expression data encoded in topological summaries, specifically PH-based PIs, along with WSI to predict resistance to FOLFOX, a common chemotherapy regimen for CRC patients. The encoding of gene expression data in a VR complex was refined using a signed-TOM instead of distance correlation measures, facilitating the extraction of higher-dimensional topological features beyond *Betti*-0. The enhanced VR complex demonstrated an improved performance of an MLP in the generation of a signed-TOM. In the realm of WSI data, the model showed a high performance in classifying cancer stages among CRC patients. The combination of CNN and MLP produced high accuracy without overfitting, even with a relatively low number of training samples. Using transfer learning with MobileNet and topology-based gene expression data encoding, *GeTopology* emerged as a robust system for various patient classification tasks. The findings outlined here not only enhance our understanding of drug resistance in CRC patients but also lay the foundation for a robust predictive model, highlighting the potential utility of *GeTopology*. This model,

as depicted in Figure 4.1, holds substantial promise to tackle the intricate challenges linked to CRC drug resistance. The noteworthy aspect of these findings lies in their potential utility, particularly in classification tasks. Identifying distinctive patterns and biomarkers associated with drug resistance, classification facilitates the development of precise treatment strategies. This potential utility is important to guide timely interventions, optimise resource allocation, and advance precision medicine in the context of CRC, thus highlighting the practical significance of these research results.

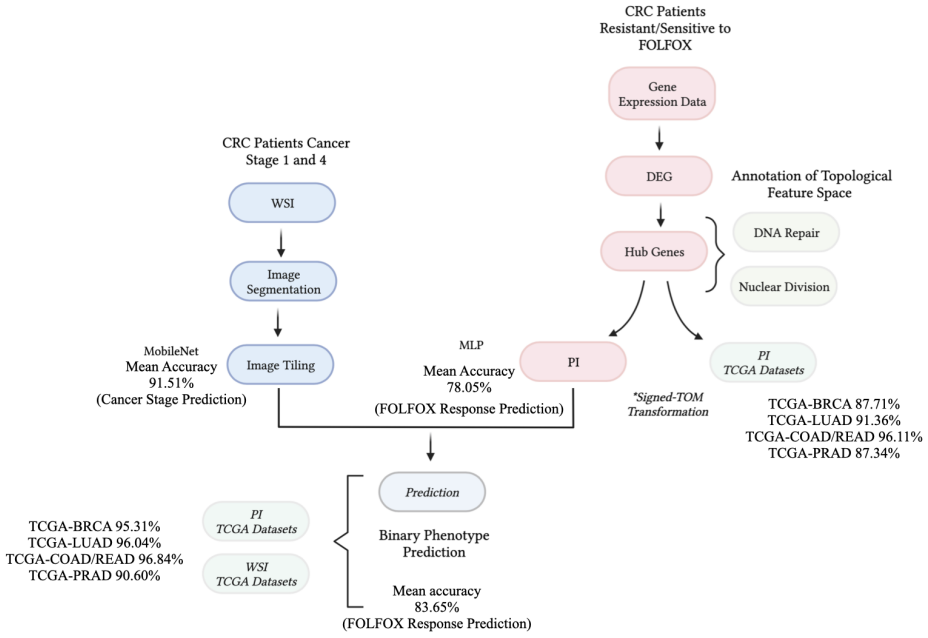


Fig. 4.1 A novel framework for colorectal cancer (CRC) analysis. Abbreviations used include WSI (Whole-Slide Image), CRC (Colorectal Cancer), DEG (Differentially Expressed Genes), PI (Persistence Image), MLP (Multilayer Perceptron), signed-TOM (Signed-Topological Overlapping Measure), TCGA (The Cancer Genome Atlas), and FOLFOX (chemotherapy regimen including folinic acid (leucovorin, "FOL"), fluorouracil (5-FU, "F"), and oxaliplatin (Eloxatin, "OX")). Notable findings encompass MobileNet achieving a 94% mean test accuracy for cancer stage prediction on WSI, identification of hub genes pivotal for DNA repair and nuclear division, and application of these hub genes in signed-TOM computation. The prediction of the FOLFOX response exhibited a 78% mean test accuracy using MLP, exceeding comparative models for PI and WSI, achieving an 83% accuracy. Specific dataset analyses demonstrated robust accuracy, emphasising the framework's efficacy. This visual representation elucidates the multidimensional successes of the study, created with Biorender.com.

4.1 Biological Insights into Differentially Expressed Genes and Hub Genes

Following the analysis and processing of QC as shown in Figures A.1, A.2 and A.3, further analysing multiple datasets provides a more comprehensive understanding of the data. This approach reveals patterns and connections that are not immediately apparent when examining individual datasets. Initially, gene expression data were subjected to differential gene expression analysis, uncovering significant changes in various biological processes (Figure A.4). This analysis highlighted a up-regulation in angiogenesis, lipid metabolism, extracellular matrix organisation, and the Wnt signaling pathway, crucial processes for cancer growth and maintenance. On the contrary, down-regulation was observed in nuclear division, antimicrobial immune responses, and lipid/fatty acid metabolism, indicating slower cell division and reduced energy consumption (Figures 3.1 and A.5). Lipid metabolism appears to be dysregulated, showing both up-regulation and down-regulation in the FOLFOX resistance samples. This phenomenon is also evident in other studies [300–302]. The detailed numbers of DEGs are provided in Table 3.1, guiding the pre-selection of gene expression matrices to reduce dimensionality. Subsequently, co-expression analysis was performed to identify hub genes, which serve as central players in gene regulatory networks.

Rather than categorising gene co-expression into binary states (connected = 1, unconnected = 0), the R programming package *WGCNA* employs a soft threshold to determine edge weights in gene co-expression networks. This approach, shown to produce less noisy networks [143], aligns the resulting co-expression network more closely with a scale-free network. The soft threshold values are summarised in Table 3.2 and illustrated in Figure A.6. The application of a scale-free topology model to GSE81653 yielded a less than satisfactory result of 77% (optimal results $\geq$ 80%) [143]. Possible data inconsistencies with a power-law distribution are considered, which can reduce the predictive power of *WGCNA* on this dataset. The organisation of the weighted adjacency matrices into gene modules and their representation as gene dendrograms (Figure 3.2) revealed distinct groups with minimal overlap between the gene modules. Analysis of hub genes in the GSE81653 dataset, with a scale-free topology fit below 80%, indicated weak links (calculated by k_{ME}). The consequences of this lead to the imposing of more stringent k_{ME} limits on this dataset, resulting in fewer hub genes, as demonstrated in Figure A.7 and mean k_{ME} shown in Table 3.3.

Analysis of four gene expression datasets highlighted the association of hub genes with DNA damage repair processes, such as binding to mismatch DNA repair complex, NER and BER (Figure A.8, 3.3, and Table 3.4). DNA repair mechanisms that address cell stress caused by DNA damage play a crucial role in reducing chemotherapy-induced damage, improving the potency of treatment [303]. Approximately 20% of the development of CRC is attributed to malfunctioning mismatch repair proteins. Patients with CRC and impaired mismatch repair systems exhibit poor overall survival and clinical outcomes [304]. Several studies have reported that the mechanisms of BER and mismatch repair are involved in resistance to FOLFOX [303–307]. The observation of elevated DNA repair mechanisms that counteract DNA damage induced by 5-FU-based therapy, such as FOLFOX, is not a novel discovery. However, these consistent findings across datasets provide valuable support for recent research efforts. The confirmation of DNA repair mechanisms in different studies has significant implications for biomarker research, shedding light on elements that contribute to FOLFOX resistance in patients with CRC. This reaffirmation of previous findings strengthens the scientific understanding of the role of DNA repair in mediating resistance to FOLFOX-based therapies in CRC. To create a clear FOLFOX response classification framework, it was essential to link the hub genes to biological functions, providing insight into the inner workings and making it easier to comprehend how classifications are made.

The acknowledgement of potential biases introduced through hub gene pre-selection in the present study is crucial for transparency and robustness. However, addressing these biases using four datasets under specific experimental conditions is a strategic approach that aims to reduce sampling biases. The use of hub genes that conform to a scale-free topology network, as highlighted in the study, carries implications for the integrity of the genetic expression network. Reference to the scale-free topology network, as proposed by Albert and Barabási (2005), emphasises the decisive impact of the hub genes on the maintenance of the structural integrity and organisation of the network [308]. This implies that the selection of hub genes is not arbitrary, but strategically aligns with the properties of the network crucial for biological relevance. The implication here is that the removal of the hub genes could alter the overall structure of the network and potentially compromise its biological significance [309]. The potential implications of addressing biases through the use of specific datasets and the emphasis on hub genes conforming to a scale-free topology network involve bolstering the reliability of the study's findings and highlighting the critical role of hub genes in maintaining the integrity and functionality of the genetic expression

network. These considerations contribute to the robustness and biological relevance of the study's conclusions.

4.2 Topological Summaries Comparisons

A graph structure or VR complex can be used to illustrate the links between the hub genes in a gene expression matrix. This can be improved by selecting a suitable correlation metric, which can then be used to form the edges of the complex [310]. Signed-TOM is one such approach that has been found to be effective in the results obtained. In TDA, the distance correlation measure is a common choice because it can measure both linear and non-linear relationships between gene pairs [270]. However, the calculation of signed-TOM takes into consideration the strength of the link between two nodes by counting the mutual neighbourhoods they share. This allows the detection of intricate interactions between genes, such as the influence of gene expression on other genes. Furthermore, the number of connections that have a positive or negative correlation with both is evaluated, rather than simply recognising the existence or lack of a relationship [251]. This provides a more precise representation of the correlation between the two nodes, rather than just the presence or absence of a link. Numerous studies on the application of signed-TOM for the analysis of co-expression networks have demonstrated the strength of signed-TOM as a representation for weighted co-expression networks [311–315].

In the present study, the effectiveness of signed-TOM as a representation of the VR complex has been demonstrated by comparing the signed-TOM and distance correlation measures. This comparison resulted in contrasting topological summaries, which are shown in Figure 3.4, which shows more topological features and higher *Betti* numbers (i.e. $Betti > 1$) in the signed-TOM constructed VR complex compared to distance correlation measures. Figure A.9, applied on independent TCGA data, also shows the same trend. The signed-TOM approach revealed a greater number of topological features in both *Betti*-1 and *Betti*-2, indicating that the orientation of the hub genes in the VR complex provided a more comprehensive description of the gene expression data. The greater the dimension of the *Betti* number (e.g. *Betti*-2, 3, 4, etc.), the more intrinsic information it contains, which could be advantageous for subsequent classification tasks [316].

Understanding the concept of *Betti* numbers can be challenging, but can be conceptualised as the count of "voids" formed by the connections between nodes or genes within a given space. This topological invariant is crucial to identifying inherent spatial structures independent of metrics [176]. Unlike general measurements such as standard correlation coefficients, *Betti* numbers are more sensitive to the specific positioning of nodes (or in this case genes) in a topological space, as illustrated by Duarte (2022) and Küster *et al* (2020) [317, 318]. This implies that even slight changes in gene location, driven by variations in gene expression values, can significantly influence *Betti* numbers.

In a more comprehensive study, Gu *et al.* (2022) highlighted the effectiveness of *Betti* number-enhanced methods, specifically *Betti*, in change point detection (CPD) within multi-agent systems [319]. CPDs denote changes or transition points, such as changes in social network behaviour indicating events of group dynamics. This example emphasises how *Betti* numbers capture information that effectively distinguishes between different states or biological phenotypes. This insightful application of *Betti* numbers in recognising transitions and changes serves as a valuable precedent for their potential utility in predictive learning tasks, such as the present study for the classification of patients with CRC based on resistance or sensitivity to FOLFOX.

4.3 Prediction Learning Comparisons

FOLFOX Resistance/Sensitive CRC Patient Classification Using Topological Summaries

MLPs are shown to be advantageous neural networks capable of identifying intricate relationships within complex datasets, particularly suitable for high-dimensional and abstract representations such as PIs [320, 321]. PIs generated from signed TOMs demonstrated improved classification performance. The MLP classification with signed TOM-generated PIs achieved favourable results, with a training mean accuracy of 80.96% and a testing mean accuracy of 78.05%. Furthermore, the AUC reached 0.85 and 0.82 for the training and testing datasets, respectively. Furthermore, essential metrics such as TPR, precision, and macro F-1 scores exceeded 78% for the test dataset. On the contrary, the MLP classification using PIs generated from distance correlation metrics showed signs of overfitting, with a training mean accuracy of 81.74% and a testing mean accuracy of 63.11% (Table 3.6). The AUC for the training data was 0.78, and for the testing data 0.67 (Figure 3.6). Furthermore, the TPR, precision, and

macro F-1 scores were approximately 65% for the test dataset. This 15% disparity in classification results between the PIs generated from signed-TOM and distance correlation metrics suggests potential overfitting. The overfitting is probably attributed to the small number of *Betti* numbers (Figure 3.4), and the small number of training samples for the MLP classifier could have contributed to the low performance metrics (Table 2.4) [322].

To further investigate this observation, independent datasets with sufficient training sample sizes were used for the classification of cancer patients and healthy individuals (Table A.1). The MLP parameters were fine-tuned for unbiased classification tasks (Table A.2). Four datasets were reviewed, including breast adenocarcinoma (BRCA), lung adenocarcinoma (LUAD), colon and rectal adenocarcinoma (COAD/READ), and prostate adenocarcinoma (PRAD). The mean training and test accuracies for these datasets were relatively high and comparable, indicating an effective generalisation of the model (Table A.2). BRCA with a mean training accuracy of 91.36% and a mean test accuracy of 87.71%, LUAD with a mean training accuracy of 93.25% and a mean test accuracy of 91.36%, COAD/READ with a mean training accuracy of 96.47% and a mean test accuracy of 96.11%, and PRAD with a mean training accuracy of 91.00% and a mean test accuracy of 87.34%. Importantly, these results were exclusive to the MLP trained and tested using signed-TOMs. Lower performance was observed for topological summaries generated using distance correlation measures, as depicted in Figure A.10. This indicates that the MLP model has learnt the patterns from the data and not artefacts, more particularly for PIs generated from signed-TOMs.

Few studies have explored the use of TDA as a data encoding method to predict phenotypes from transcriptomic data. Recent studies by Shapanis *et al.* (2023) and Wamil *et al.* (2023) have illustrated the versatility of TDA in unraveling complex patterns within heterogeneous disease populations, contributing to improved risk prediction and understanding of clinical outcomes [175, 323]. The foundational work in this line of research was carried out by Mandal *et al.* (2020), demonstrating comparable results. Their use of a vectorised form of the persistent diagram resulted in a prediction performance that exceeded 86% [324]. Specifically, their task aimed to classify Parkinson's patients and healthy individuals using RNA-Seq data. Taking into account the more challenging task of classifying CRC patients as resistant or sensitive to FOLFOX treatment, our mean accuracy of 78.05% is 8% less than that achieved by Mandal *et al.* (2020). However, our model, which performs a similar task of classifying

disease and non-disease, achieved a classification mean accuracy of 87.71% (BRCA), 91.36% (LUAD), 96.11% (COAD/READ), and 87.34% (PRAD). This indicates that our approach is comparable to Mandal *et al.* (2020) but suggests an improvement through the implementation of singed-TOMs in the pre-processing step compared to the use of distance correlation measures employed by Mandal *et al.* (2020). Collectively, these findings highlight the growing recognition of TDA as a powerful tool for decoding intricate biological patterns and predicting phenotypic outcomes from transcriptomic data.

The noteworthy discovery in this work, which illustrates improved accuracy in predicting phenotypes through the application of signed-TOM to gene expression data, extends beyond merely predicting resistance to FOLFOX. This enhancement is attributed to the consideration of direction and connection strength in relation to neighbouring connections of topological features, particularly beneficial in persistence-based filtration. Importantly, this novel connection was not discovered in previous research, highlighting the uniqueness of this approach. The implications of this finding are substantial. Firstly, it suggests a more subtle understanding of the relationship between gene expression and phenotypic outcomes, indicating that considering the direction and strength of the connection in relation to neighbouring connections of topological features adds valuable information for accurate predictions. This insight could potentially reshape how researchers approach the analysis of gene expression data in the context of phenotypic predictions.

In the context of the TCGA-COAD/READ cohort, the goal to improve predictive accuracy involves the integration of an additional data type, WSI, into the prediction task. WSI pre-processing, including crucial steps such as image segmentation and tiling, is performed to effectively prepare the images for subsequent analyses. This integration of morphological information, alongside gene expression data, results in a more comprehensive dataset, providing a multi-dimensional perspective that contributes to the refinement of the predictive models. The inclusion of various data types, as demonstrated by thorough pre-processing of WSI, has promising implications for future research, particularly in advancing the robustness of predictive models for cancer phenotypes and treatment responses.

Cancer Classification using WSI

Predicting drug resistance in cancer cells based on visible features of WSI poses a challenging task. The intricate molecular and genetic processes involved in drug resistance may not be easily discernible through CNNs, as demonstrated by Chiu *et al.* (2019) [325]. However, recent advances in WSI analysis have demonstrated the ability to systematically identify and interpret the relationship between patient histological patterns, multiomics, and clinical profiles. Unexpected correlations have been observed between changes in histopathological images and clinically actionable molecular variations [326–329]. For example, Kather *et al.* trained a CNN model to predict genome microsatellite instability in gastrointestinal cancer from H&E histology, achieving a patient-level AUC of 0.84 [330]. Coudray *et al.* successfully trained a DL network that accurately predicted six out of the ten most commonly mutated genes in the pathology images of TCGA-LUAD, with AUCs ranging from 0.733 to 0.856 [331]. Subsequently, Kather *et al.* demonstrated the ability of DL to predict point mutations, molecular tumour subtypes, and immune-related gene expression signatures directly from H&E images across multiple types of cancer [332]. Fu *et al.* used transfer learning, correlating the characteristics of histopathological patterns with genomic, transcriptomic, and survival data in 28 types of cancer [333]. It should be noted that these recent studies use complex DL/ML approaches, often incorporating transfer learning with deep CNNs and frequently integrating multiple DL/ML models. The wealth of data contained in WSI can be used to uncover the intratumor heterogeneity, which can then be used to better understand complex biological processes such as drug resistance [331–333].

Despite the complexity of molecular interactions within cancer cells, accurately predicting drug resistance solely on the basis of WSI with simple CNN remains a challenge. However, certain physical features, such as cell size, shape and nucleus-cytoplasm ratio, can still be used to identify drug resistant cells considering alterations in cell characteristics, including adhesion, junctions, and contacts [334, 335]. These features can also provide valuable data on the size, shape and arrangement of cancer cells, helping to determine the stage of cancer [336–338]. Furthermore, investigating the relationship between cancer stage, drug resistance, and survival is best done by focussing on the cancer stage rather than evaluating resistance to FOLFOX. Cancer stage provides a more comprehensive measure of cancer progression, considering factors such as tumour size, extent of spread, and other variables that influence prognosis. Studies have shown that advanced cancers exhibit a higher degree of variation in cell size and shape, necrosis, and mitotic activity [339–341]. Evaluating cancer stage rather

than solely focussing on FOLFOX resistance allows a more holistic understanding of cancer progression and its impact on patient outcomes. In particular, CNNs have been proven to be effective in forecasting cancer stages by capturing spatial information from medical images. Wang *et al.* (2022) demonstrated a 90% precision in estimating the stage of CRC from computed tomography (CT) scans using a CNN in a transfer learning context, which was an improvement over conventional CNN trained from scratch.[342]. Similarly, Zhang *et al.* (2020) reported 85% accuracy in identifying stages of breast cancer with a CNN using magnetic resonance imaging (MRI) [343]. The research presented supports the efficacy of CNNs with regards to transfer learning in determining the stage of cancer. Furthermore, MobileNets have been shown to be more effective in complex biomedical image classification tasks, such as WSIs [344–347].

The present study demonstrated that MobileNets had remarkable results, exceeding even the MLP classification of PIs. The MobileNet model for cancer stage prediction exhibited an average training accuracy of 94.58% and an average test accuracy of 91.51%, with AUC of 0.96 and 0.95, respectively (Table 3.7 and Figure 3.5). The implications of the performance of the present study model and its effective approach in predicting cancer stages are significant. Some notable MobileNet implementations with high prediction accuracy, comparable to our framework, include Kumar *et al.* (2021), who conducted a binary classification task for benign and malignant phenotypes in breast cancer using WSI and reported a mean accuracy of 89% [348]. Furthermore, Laxmisagar *et al.* (2022) used a MobileNet to classify four breast cancer subtypes with a resulting mean accuracy of 87.5% [349]. Lastly, Sharma *et al.* (2023) performed a binary classification task for the detection of invasive ductal carcinoma in breast cancer and reported an 86% mean accuracy using MobileNet [350]. These findings suggest that the present study offers a more reliable and accurate method to perform classification tasks. The success can be attributed to data quality and meticulous pre-processing emphasises the importance of these factors in the development of robust predictive models for cancer staging. This information could influence future research and applications in medical imaging and cancer diagnostics.

Although WSI may not fully represent genetic and molecular differences within a tumour, efforts are underway to link gene alterations with morphological patterns in WSI [331, 329, 351]. Although the application of CNNs for drug resistance prediction may become more feasible as this research area progresses, certain considerations must be addressed. Access to large, labelled datasets is crucial for training and validating

ML/DL models. Limited and non-comprehensive datasets can make models susceptible to overfitting [352]. Understanding the specific histopathological features associated with predicting drug response is essential for the successful clinical use of CNNs. Furthermore, considering the evolving nature of drug resistance over time, longitudinal data becomes necessary, which is not always available in WSI datasets [353]. However, combining histopathological information with other types of data, such as genomic or transcriptomic data, can enhance the accuracy of these models, as demonstrated in the present study [354–356].

CRC Patient Classification using Topological Summaries and WSI

The application of DL to bimodal data in the biological field has seen a marked increase in recent years, especially with respect to histopathological data such as WSI, proteomic, and genomic data [357–359]. This trend has resulted in improved performance optimisation when dealing with multi-omics datasets, leading to the development of reliable ensemble methods prediction models that find applications in precision medicine. Consequently, the integration of two distinct types of data in DL has shown the potential to significantly enhance model predictions, particularly in scenarios with limited samples. The joining of two data types serves to mitigate the risk of overfitting, enabling the model with diverse sources of information, and, consequently, yielding more reliable and precise outcomes. Bimodal learning not only contributes to improved accuracy but also decreases the amount of data required for training, rendering the model computationally more efficient.

The *GeTopology* framework demonstrates notable improvements in mean accuracy during the prediction of FOLFOX resistance in CRC patients, using a combination of PIs and CNN classification of WSI cancer stages. The mean accuracy rates increased to 84.32% in training and 83. 65% in the tests, demonstrating improvements compared to the use of PIs with a mean training accuracy of 80.96% and a mean testing accuracy of 78.05 %. The integration of CNN classification contributed to this performance boost. A comparative analysis between *GeTopology* and *NSCLC* for FOLFOX resistance prediction in CRC patients revealed marginal differences, with *GeTopology* exhibiting an approximate 2% performance increase (Figure 3.7 and Table 3.8). Further evaluations on independent cancer datasets, focussing on the classification of cancer patients and healthy individuals, consistently showed that *GeTopology* outperformed *NSCLC* across multiple datasets (BRCA, LUAD, COAD/READ, and PRAD). For example, in the

BRCA dataset, *GeTopology* achieved a mean test accuracy of 95.31%, exhibiting a 9% improvement over *NSCLC's* 84%. In general *GeTopology* achieved higher mean test accuracies in all datasets, indicating its robust performance compared to *NSCLC* (Table A.3 and Figure A.11).

Improvements in model performance, particularly in predicting FOLFOX resistance, can be attributed to an increase in the number of training samples and the use of matched patient samples comprising PI (gene expression data) and WSI. *GeTopology* stands out by incorporating TDA, contributing to its superior model performance. TDA, specifically the integration of *Betti* numbers, provides a subtle understanding of the structural characteristics of gene expression patterns, offering a unique perspective on the tumour microenvironment.

Topology, as applied in TDA, proves to be a valuable tool for encoding complex structures and simplifying information representation. Its reliability and resilience to outliers make it particularly useful for uncovering patterns and connections within gene expression data. Studies, such as those of Ray *et al.* (2010), Nicolau *et al.* (2011), and Asaad *et al.* (2022), have demonstrated the effectiveness of these topological summaries in recognising gene sets related to specific biological processes or diseases, including Alzheimer's and breast cancer [360, 361, 173]. Furthermore, other studies have used topological summaries to improve predictive analysis, as evident in the works of Townsend *et al.* (2020), Dey *et al.* (2021), Pun *et al.* (2020), Mandal *et al.* (2020), Carriere *et al.* (2020), Meng *et al.* (2020), Vandaele *et al.* (2023), Liu *et al.* (2021), Hofer *et al.* (2017) and Hirata *et al.* (2020) [362–364, 324, 365, 366, 176, 367–369]. The use of topological summaries, including *Betti* numbers, offers a comprehensive perspective on gene expression, facilitating a deeper understanding of the intricate relationships between genes and their expression levels.

The potential implications of this approach extend to patient stratification in precision medicine. When topology is applied in TDA, a more detailed exploration of gene expression data becomes possible, enabling the identification of intricate patterns that are often overlooked by traditional methods. This is particularly relevant for patient classification, where recognising gene sets associated with specific biological processes or diseases is crucial for tailoring treatments. The use of topological summaries in predictive analysis offers a promising avenue to enhance the accuracy and reliability of models, affecting various applications, including precision medicine. Accurate

predictions based on gene expression and WSI data are essential to guide treatment decisions, making topological approaches invaluable. In conclusion, the incorporation of topology in TDA, especially through PIs, has demonstrated notable success in patient stratification and various aspects of genomic research. From discovering disease-related gene sets to improving predictive analysis, topological approaches offer a comprehensive framework to advance the understanding of gene expression and its implications for diseases and precision medicine. While promising, it is essential to acknowledge the study's limitations to provide a balanced perspective on its findings and implications.

4.4 Limitations of the Present Study

Data Availability

The study encountered significant challenges in accessing RNA-Seq data for patients exhibiting FOLFOX resistance and sensitivity. These challenges stemmed from limitations in sample numbers and variations in experimental conditions. Consequently, the study had to rely heavily on microarray data to navigate these data availability constraints. The restricted availability of RNA-Seq data, particularly for FOLFOX resistance datasets, can be attributed to the competitive nature of drug resistance research. Researchers may hesitate to share extensive datasets due to concerns about the priority and credit of discoveries, thereby limiting data sharing in this competitive environment [370]. In addition, the generation of comprehensive drug resistance datasets is a complex and resource-intensive process that requires substantial time, effort, and financial resources to conduct clinical studies, laboratory experiments, and collect patient data. The challenges associated with the creation of such datasets often result in their limited size, especially in the early stages of research [371]. These limitations highlight the intricacies associated with data access in the study, highlighting the difficulty and challenges associated with interpreting and generalising the findings.

Computational Cost

In the construction of the VR complex, an optimal strategy would have involved the inclusion of all genes. However, the computational challenge has arisen from the complexity of computing pairwise gene correlations within the complete set of 19,116 genes, which proved to be nearly insurmountable. To address potential biases related to

pre-selection, the study opted for a compromise by limiting the inclusion of all available genes. This computational constraint highlights a limitation in terms of computational cost, emphasising the need for pragmatic approaches in large-scale genetic analyses.

The TDA for PH computation and the use of MobileNet as a deep neural network imposed substantial demands on computational resources. The study relied on a high performance system equipped with 32 gigabytes of random access memory (RAM), an Intel i7 8th generation processor, and a GeForce RTX2060 graphics processing unit (GPU). Although crucial for performing computational tasks, this powerful system incurred higher computational expenses compared to the *NSCLC* framework, which demonstrated efficient operation with less RAM. This computational limitation emphasises the resource-intensive nature of employing advanced methodologies in genetic analyses. The present study emphasises the challenges associated with computational costs and the need to consider the scalability of computational infrastructure in similar studies.

4.5 Conclusions

The present study aimed to advance the understanding of CRC through the innovative integration of PH-based PIs and WSIs. The application of signed-TOM to construct the VR complex within PH captured various topological features, including high-dimensional *Betti*-1 and *Betti*-2, within genetic expression data, with a particular emphasis on hub genes. The critical role of hub genes as integral components within gene networks, which govern biological functions related to DNA repair and nuclear division, was highlighted. At the same time, WSIs provided intricate microscopic insights into the tissue anatomy crucial for delineating cancer stages. The integration of genetic expression data represented by PIs with morphological information from WSIs resulted in a comprehensive dataset that facilitates intricate predictions. This fusion capitalised on both molecular and spatial characteristics. The core contribution of the study lies in decoding the intricate interplay between genetic and morphological aspects, which offers more accurate predictions and a deeper understanding of the tumour microenvironment. Implications of the findings of the present study include that the use of signed-TOM for TDA computation demonstrated the enhancement of capturing higher-dimensional topological features, adding a novel dimension to the analytical approach. The accurate prediction of cancer stages using WSI further

validated the clinical relevance of the integrative methodology presented.

Importantly, the DL model that uses two modalities of data, which incorporates both genetic and morphological data, has critical implications in the context of cancer biology. By leveraging the synergy between these modalities, the present study offers a more holistic understanding of the response to FOLFOX and the staging of cancer, with promising prospects for precise therapeutic interventions. The potential clinical applications of this integrated approach include more targeted and effective treatments that improve patient outcomes in CRC. For training, a large number of matched gene expression samples and corresponding WSI are required to build robust predictive models. This entails collecting and preprocessing a diverse dataset that encompasses a wide range of genetic profiles and histopathological characteristics. Once the models are trained using this comprehensive dataset, the deployment phase involves processing individual patient data. For each patient, a single mRNA sample, representing gene expression, and a WSI of the tumour biopsy, which has been cross-sectioned, are inputted into the trained models. These inputs are then analysed by the models to predict various phenotypic characteristics, such as disease prognosis, treatment response, or molecular subtypes. This approach allows for personalised and precise predictions based on the specific genetic and histopathological features of each patient's tumour sample, aiding clinicians in making informed treatment decisions tailored to individual patient needs.

In summary, the implications of this finding extend beyond the specific improvement in the prediction of FOLFOX resistance. These findings suggest a novel avenue for refining predictive models using signed-TOM to identify topological features and integrate diverse data types. This advancement in methodology holds promise for improving the prediction and understanding of complex phenotypes in the context of cancer biology. For a comprehensive description of the framework's implementation, please refer to Section B. In conclusion, this study makes substantial contributions by advancing genomic and histopathological methodologies, providing novel insights into the classification of complex diseases, particularly CRC. The integration of topological and spatial features using PIs and WSI respectively represents a promising avenue for future research and applications in precision medicine for CRC patients.

4.6 Recommendations for Future Enhancements in the Prediction Models

Parameter Customisation and User Interface

To enhance user accessibility, the code will undergo updates, facilitating seamless parameter customisation. This improvement aims to address a diverse range of researchers, including those with varying levels of computational expertise. For example, adjustable parameters such as training, test and validation data sets, learning rate, and algorithm specifications for TDA calculations are currently set as default values. The upcoming code revisions will allow users to easily customise these parameters according to their specific research requirements. However, it is essential to emphasise that any changes to these parameters should be approached with caution, as they can significantly impact the results of the analysis.

Mitigating Bias in TDA Computations

To circumvent the need for pre-selection and mitigate potential biases, the inclusion of all genetic data becomes imperative. Performing this involves the refinement and optimisation of algorithms, particularly signed-TOM, to generate a matrix of dimensions 19,116 $\times$ 19,116 for the construction of the VR complex. However, the computational challenges inherent in PH calculations arise from the extensive matrix, which prompts the application of a basic collapse technique to streamline VR complexes through edge collapse. This process effectively reduces the dimension of VR complexes while preserving the optical topology of the complex while effectively reducing noise, thus serving as a crucial mitigation step to eliminate biases in TDA computations [372].

Topology-based Algorithm Refinement

In refining the construction of the VR complex, the incorporation of various edge weighting strategies proves vital. Adopting different edge weighting approaches becomes crucial to mitigate the impact of biased or noisy edges on the resulting topology. A recent technique demonstrated by Mandal *et al.* (2020) demonstrates the effectiveness of assigning lower weights to such edges, providing valuable information [275, 324]. These strategies are recommended for consideration in the present study to improve

the accuracy and robustness of the VR complex.

Interpretation and Validation

Collaboration with biologists and domain experts is imperative for the validation and accurate interpretation of the topological features derived from TDA. Working closely with experts in the biological or clinical field ensures that the TDA results align seamlessly with the biological context, minimising the risk of biases or misinterpretations. The validation of TDA results remains of critical importance, and a meticulous comparison with known biological processes or experimental validations is required to confirm their authenticity. In the scope of this study, the use of TDA to identify *Betti* numbers, representing topological features known as 'voids', adds a unique dimension. These *Betti* numbers encapsulate sets of informative genes from a topological perspective, offering distinctive insights into their spatial relationships within biological networks.

To deepen the understanding of these gene sets, particularly their biological functions, it is recommended to conduct *in vitro* and *in vivo* studies. Investigating the functions and interactions of these genes in disease contexts can provide valuable information about their roles and potential implications in pathological processes. This comprehensive approach, seamlessly combining TDA with subsequent biological investigations, enhances the robustness and translational relevance of the study findings. Such collaborative efforts not only bolster the validity of the results of TDA but also highlight its potential as a groundbreaking approach to identify novel biomarkers and unravel complex disease outcomes.

References

[1] John S Bertram. The molecular biology of cancer. *Molecular Aspects of Medicine*, 21(6):167–223, 2000.

[2] John J Tyson. Modeling the cell division cycle: cdc2 and cyclin interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 88(16):7328–7332, 1991.

[3] Valentina Bollati and Andrea Baccarelli. Environmental epigenetics. *Heredity*, 105(1):105–112, 2010.

[4] Wynand P Roos, Adam D Thomas, and Bernd Kaina. Dna damage and the balance between survival and death in cancer biology. *Nature Reviews Cancer*, 16(1):20–33, 2016.

[5] Kaleigh Fernald and Manabu Kurokawa. Evading apoptosis in cancer. *Trends in Cell Bbiology*, 23(12):620–633, 2013.

[6] Rebecca SY Wong. Apoptosis in cancer: From pathogenesis to treatment. *Journal of Experimental & Clinical Cancer Research*, 30(1):87, 2011.

[7] Yuri Lazebnik. What are the hallmarks of cancer? *Nature Reviews Cancer*, 10 (4):232–233, 2010.

[8] Franziska Michor, Yoh Iwasa, and Martin A Nowak. Dynamics of cancer progression. *Nature Reviews Cancer*, 4(3):197–205, 2004.

[9] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011.

[10] Yousef Ahmed Fouad and Carmen Aanei. Revisiting the hallmarks of cancer. *American Journal of Cancer Research*, 7(5):1016, 2017.

[11] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100 (1):57–70, 2000.

[12] Raphael Trenker and Natalia Jura. Receptor tyrosine kinase activation: From the ligand perspective. *Current Opinion in Cell Biology*, 63:174–185, 2020.

[13] Yan-Jun Guo, Wei-Wei Pan, Sheng-Bing Liu, Zhong-Fei Shen, Ying Xu, and Ling-Ling Hu. Erk/mapk signalling pathway and tumorigenesis. *Experimental and Therapeutic Medicine*, 19(3):1997–2007, 2020.

[14] Roberta Lugano, Mohanraj Ramachandran, and Anna Dimberg. Tumor angiogenesis: Causes, consequences, challenges and opportunities. *Cellular and Molecular Life Sciences*, 77:1745–1770, 2020.

[15] Mehdi Rajabi and Shaker A Mousa. The role of angiogenesis in cancer treatment. *Biomedicines*, 5(2):34, 2017.

[16] Lena Claesson-Welsh and Michael Welsh. Vegfa and tumour angiogenesis. *Journal of Internal Medicine*, 273(2):114–127, 2013.

[17] Nathaniel J Robinson and William P Schiemann. Telomerase in cancer: Function, regulation, and clinical translation. *Cancers*, 14(3):808, 2022.

[18] Catherine Cifuentes-Rojas and Dorothy E Shippen. Telomerase regulation. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 730 (1-2):20–27, 2012.

[19] Tomasz Trybek, Artur Kowalik, Stanisław Góźdź, and Aldona Kowalska. Telomeres and telomerase in oncogenesis. *Oncology Letters*, 20(2):1015–1027, 2020.

[20] Mark S D'arcy. Cell death: A review of the major forms of apoptosis, necrosis and autophagy. *Cell Biology International*, 43(6):582–592, 2019.

[21] Puneet Singh and Bora Lim. Targeting apoptosis in cancer. *Current Oncology Reports*, 24(3):273–284, 2022.

[22] Benedito A Carneiro and Wafik S El-Deiry. Targeting apoptosis in cancer therapy. *Nature Reviews Clinical Oncology*, 17(7):395–417, 2020.

[23] Karin Hientz, André Mohr, Dipita Bhakta-Guha, and Thomas Efferth. The role of p53 in cancer drug resistance and targeted chemotherapy. *Oncotarget*, 8(5): 8921, 2017.

[24] Christine L Chaffer and Robert A Weinberg. A perspective on cancer cell metastasis. *Science*, 331(6024):1559–1564, 2011.

[25] Bethany N Smith and Neil A Bhowmick. Role of emt in metastasis and therapy resistance. *Journal of Clinical Medicine*, 5(2):17, 2016.

[26] Veena Padmanaban, Ilona Krol, Yasir Suhail, Barbara M Szczerba, Nicola Aceto, Joel S Bader, and Andrew J Ewald. E-cadherin is required for metastasis in multiple models of breast cancer. *Nature*, 573(7774):439–444, 2019.

[27] Saima Usman, Naushin H Waseem, Thuan Khanh Ngoc Nguyen, Sahar Mohsin, Ahmad Jamal, Muy-Teck Teh, and Ahmad Waseem. Vimentin is at the heart of epithelial mesenchymal transition (emt) mediated metastasis. *Cancers*, 13(19): 4985, 2021.

[28] Mathias Heikenwalder and Anna Lorentzen. The role of polarisation of circulating tumour cells in cancer metastasis. *Cellular and Molecular Life Sciences*, 76:3765–3781, 2019.

[29] Marion Piñeros, D Maxwell Parkin, Kevin Ward, Eric Chokunonga, Morten Ervik, Helen Farrugia, Mary Gospodarowicz, Brian O'Sullivan, Isabelle Soerjomataram, and Rajaraman Swaminathan. Essential tnm: A registry tool to reduce gaps in cancer staging information. *The Lancet Oncology*, 20(2):e103–e111, 2019.

[30] Hideki Ueno, Hidetaka Mochizuki, Yoshito Akagi, Takaya Kusumi, Kazutaka Yamada, Masahiro Ikegami, Hiroshi Kawachi, Shingo Kameoka, Yasuo Ohkura, and Tadahiko Masaki. Optimal colorectal cancer staging criteria in tnm classification. *Journal of Clinical Oncology*, 30(13):1519–1526, 2012.

[31] Yun Shin Chun, Timothy M Pawlik, and Jean-Nicolas Vauthey. 8th edition of the ajcc cancer staging manual: Pancreas and hepatobiliary cancers. *Annals of Surgical Oncology*, 25:845–847, 2018.

[32] Guillaume Onyeaghala, Anna K Lintelmann, Corrine E Joshu, Pamela L Lutsey, Aaron R Folsom, Kimberly Robien, Elizabeth A Platz, and Anna E Prizment. Adherence to the world cancer research fund/american institute for cancer research cancer prevention guidelines and colorectal cancer incidence among african americans and whites: The atherosclerosis risk in communities study. *Cancer*, 126(5):1041–1050, 2020.

[33] DM Gomez Zubieta and G Kwakye. Colorectal cancer: A review of epidemiological trends and screening guidelines. *South African General Practitioners Journal*, 4 (1):10–12, 2023.

[34] Camilla Mattiuzzi, Fabian Sanchis-Gomar, and Giuseppe Lippi. Concise update on colorectal cancer epidemiology. *Annals of Translational Medicine*, 7(21), 2019.

[35] Md Sanower Hossain, Hidayah Karuniawati, Ammar Abdulrahman Jairoun, Zannat Urbi, Der Jiun Ooi, Akbar John, Ya Chee Lim, KM Kaderi Kibria, AKM Mohiuddin, and Long Chiau Ming. Colorectal cancer: A review of carcinogenesis, global epidemiology, current challenges, risk factors, preventive and treatment strategies. *Cancers*, 14(7):1732, 2022.

[36] Firas Baidoun, Kholoud Elshiwy, Yasmine Elkeraie, Zahi Merjaneh, George Khoudari, Muhammad T Sarmini, Mohamed Gad, Muneer Al-Husseini, and Anas Saad. Colorectal cancer epidemiology: recent trends and impact on outcomes. *Current Drug Targets*, 22(9):998–1009, 2021.

[37] Allison Drosdowsky, Karen E Lamb, Rebecca J Bergin, Lucy Boyd, Kristi Milley, Maarten J IJzerman, and Jon D Emery. A systematic review of methodological considerations in time to diagnosis and treatment in colorectal cancer research. *Cancer Epidemiology*, 83:102323, 2023.

[38] Sigurdis Haraldsdottir, Hulda M Einarsdottir, Agnes Smaradottir, Adalsteinn Gunnlaugsson, and Thorvardur R Halfdanarson. Colorectal cancer-review. *Laeknabladid*, 100(2):75–82, 2014.

[39] Leah H Biller and Deborah Schrag. Diagnosis and treatment of metastatic colorectal cancer: A review. *Journal of the American Medical Association*, 325 (7):669–685, 2021.

[40] Md Sanower Hossain, Hidayah Karuniawati, Ammar Abdulrahman Jairoun, Zannat Urbi, Der Jiun Ooi, Akbar John, Ya Chee Lim, KM Kaderi Kibria, AKM Mohiuddin, and Long Chiau Ming. Colorectal cancer: A review of carcinogenesis, global epidemiology, current challenges, risk factors, preventive and treatment strategies. *Cancers*, 14(7):1732, 2022.

[41] Rachel M McQuade, Vanesa Stojanovska, Joel C Bornstein, and Kulmira Nurgali. Colorectal cancer chemotherapy: The evolution of treatment and new approaches. *Current Medicinal Chemistry*, 24(15):1537–1557, 2017.

[42] Pasqualino Favoriti, Gabriele Carbone, Marco Greco, Felice Pirozzi, Raffaele Emmanuele Maria Pirozzi, and Francesco Corcione. Worldwide burden of colorectal cancer: A review. *Updates in Surgery*, 68:7–11, 2016.

[43] So Hyun Kang, Sa-Hong Min, Jin Won Kim, Eunju Lee, Sang Woo Park, Sangjun Lee, Hyeon Jeong Oh, Young Suk Park, Yoon Jin Lee, and Ji-Won Kim. Safety and efficacy of intraperitoneal paclitaxel plus intravenous fluorouracil, leucovorin, and oxaliplatin (folfox) for gastric cancer with peritoneal metastasis. *Annals of Surgical Oncology*, 29(8):5084–5091, 2022.

[44] Paula M De Angelis, Debbie H Svendsrud, Katherine L Kravik, and Trond Stokke. Cellular response to 5-fluorouracil (5-fu) in 5-fu-resistant colon cancer cell lines during treatment and recovery. *Molecular Cancer*, 5(20), 2006.

[45] Xingxing Huang, Kun Ke, Weiwei Jin, Qianru Zhu, Qicong Zhu, Ruyi Mei, Ruonan Zhang, Shuxian Yu, Lan Shou, and Xueni Sun. Identification of genes related to 5-fluorouracil based chemotherapy for colorectal cancer. *Frontiers in Immunology*, 13:887048, 2022.

[46] Chengren Zhang, Lili Liu, Yaochun Lv, Jingjing Li, Cong Cao, Jiyong Lu, Shuai Wang, Binbin Du, and Xiongfei Yang. Chemotherapy plus panitumumab/cetuximab versus chemotherapy plus bevacizumab in wild-type kras/ras. *Expert Review of Anticancer Therapy*, 22(12):1333–1347, 2022.

[47] İlkay Gültürk and Şahingöz Erdal, Gülçin and Özmen, Aykut and Yildiz Tacar, Seher and Yilmaz, Mesut and Turalchat, Deniz. With metastatic colorectal cancer: A single center experience use of monoclonal antibodies (bevacizumab, cetuximab, and panitumumab) in patients. *Journal of Experimental & Clinical Medicine/Deneysel ve Klinik Tip Dergisi*, 39(4), 2022.

[48] Shashank J Patel, Neville E Sanjana, Rigel J Kishton, Arash Eidizadeh, Suman K Vodnala, Maggie Cam, Jared J Gartner, Li Jia, Seth M Steinberg, and Tori N Yamamoto. Identification of essential genes for cancer immunotherapy. *Nature*, 548(7669):537–542, 2017.

[49] Fabrice Yves Ndjana Lessomo, Zhiquan Wang, and Chishimba Mukuka. Comparative cardiotoxicity risk of pembrolizumab versus nivolumab in cancer patients undergoing immune checkpoint inhibitor therapy: A meta-analysis. *Frontiers in Oncology*, 13:1080998, 2023.

[50] Tanweer Haider, Vikas Pandey, Nagma Banjare, Prem N Gupta, and Vandana Soni. Drug resistance in cancer: Mechanisms and tackling strategies. *Pharmacological Reports*, 72(5):1125–1151, 2020.

[51] José Rueff and António Sebastião Rodrigues. Cancer drug resistance: A brief overview from a genetic viewpoint. In *Cancer Drug Resistance*, volume 1395 of *Methods in Molecular Biology*, pages 1–18. Humana Press, New York, NY, 2016.

[52] Lan Thi Hanh Phi, Ita Novita Sari, Ying-Gui Yang, Sang-Hyun Lee, Nayoung Jun, Kwang Seock Kim, Yun Kyung Lee, and Hyog Young Kwon. Cancer stem cells (cscs) in drug resistance and their therapeutic implications in cancer treatment. *Stem Cells International*, 18:5416923, 2018.

[53] Hamed Rezayatmand, Mahboobeh Razmkhah, and Iman Razeghian-Jahromi. Drug resistance in cancer therapy: The pandora's box of cancer stem cells. *Stem Cell Research & Therapy*, 13(1):181, 2022.

[54] Mohamed R Abdelaal and Hesham Haffez. The potential roles of retinoids in combating drug resistance in cancer: implications of atp-binding cassette (abc) transporters. *Open Biology*, 12(6):220001, 2022.

[55] Anupama Modi, Dipayan Roy, Shailja Sharma, Jeewan Ram Vishnoi, Puneet Pareek, Poonam Elhence, Praveen Sharma, and Purvi Purohit. Abc transporters in breast cancer: Their roles in multidrug resistance and beyond. *Journal of Drug Targeting*, 30(9):927–947, 2022.

[56] Lindsey Carlsen and Wafik S El-Deiry. Differential p53-mediated cellular responses to dna-damaging therapeutic agents. *International Journal of Molecular Sciences*, 22(21):11828, 2021.

[57] Zhaohui Jin and Frank A Sinicrope. Mismatch repair-deficient colorectal cancer: Building on checkpoint blockade. *Journal of Clinical Oncology*, 40(24):2735–2750, 2022.

[58] Jicheng Yue and José M López. Understanding mapk signaling pathways in apoptosis. *International Journal of Molecular Sciences*, 21(7):2346, 2020.

[59] Douglas R Green and Fabien Llambi. Cell death signaling. *Cold Spring Harbor perspectives in biology*, 7(12):a006080, 2015.

[60] Yi Qin, Milad Ashrafizadeh, Vera Mongiardini, Benedetto Grimaldi, Francesco Crea, Katja Rietdorf, and Győrffy, Balázs and Klionsky, Daniel J and Ren, Jun and Zhang, Wei. Autophagy and cancer drug resistance in dialogue: Pre-clinical and clinical evidence. *Cancer Letters*, 570:216307, 2023.

[61] Haolan Wang, Ming Guo, Hudie Wei, and Yongheng Chen. Targeting p53 pathways: Mechanisms, structures, and advances in therapy. *Signal Transduction and Targeted Therapy*, 8(1):92, 2023.

[62] Mahbuba Rahman and Mohammad Rubayet Hasan. Cancer metabolism and drug resistance. *Metabolites*, 5(4):571–600, 2015.

[63] Qianyu Wang, Xiaofei Shen, Gang Chen, and Junfeng Du. Drug resistance in colorectal cancer: From mechanism to clinic. *Cancers*, 14(12):2928, 2022.

[64] Elizabeth Varghese, Samson Mathews Samuel, and Líšková, Alena and Samec, Marek and Kubatka, Peter and Büsselberg, Dietrich. Targeting glucose metabolism to overcome resistance to anticancer chemotherapy in breast cancer. *Cancers*, 12(8):2252, 2020.

[65] Ruixue Yang, Mei Yi, and Bo Xiang. Novel insights on lipid metabolism alterations in drug resistance in cancer. *Frontiers in Cell and Developmental Biology*, 10:875318, 2022.

[66] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[67] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, and Ariel Gershman. The complete sequence of a human genome. *Science*, 376(6588): 44–53, 2022.

[68] Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Molecular Cell*, 58(4):586–597, 2015.

[69] Charlie C Xiang and Yidong Chen. cdna microarray technology and its applications. *Biotechnology Advances*, 18(1):35–46, 2000.

[70] Kirk J Mantione, Richard M Kream, Hana Kuzelova, Radek Ptacek, Jiri Raboch, Joshua M Samuel, and George B Stefano. Comparing bioinformatic gene expression profiling methods: Microarray and rna-seq. *Medical Science Monitor Basic Research*, 20:138, 2014.

[71] Roger Bumgarner. Overview of dna microarrays: types, applications, and their future. *Current Protocols in Molecular Biology*, 101(1):22–1, 2013.

[72] Armin Ehrenreich. Dna microarray technology for the microbiologist: an overview. *Applied Microbiology and Biotechnology*, 73:255–273, 2006.

[73] Danh V Nguyen, A Bulak Arpat, Naisyin Wang, and Raymond J Carroll. Dna microarray experiments: Biological and technological aspects. *Biometrics*, 58(4): 701–717, 2002.

[74] James M Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing dna. *Genomics*, 107(1):1–8, 2016.

[75] Uwe Kühn and Elmar Wahle. Structure and function of poly (a) binding proteins. *Biochimica Et Biophysica Acta - Gene Structure and Expression*, 1678(2-3):67–84, 2004.

[76] Mingye Hong, Shuang Tao, Ling Zhang, Li-Ting Diao, Xuanmei Huang, Shaohui Huang, Shu-Juan Xie, Zhen-Dong Xiao, and Hua Zhang. Rna sequencing: New technologies and applications in cancer research. *Journal of Hematology & Oncology*, 13(166):1–16, 2020.

[77] Jeffrey J Werner, Dennis Zhou, J Gregory Caporaso, Rob Knight, and Largus T Angenent. Comparison of illumina paired-end and single-direction sequencing for microbial 16s rrna gene amplicon surveys. *The ISME Journal*, 6(7):1273–1276, 2012.

[78] Karl V Voelkerding, Shale A Dames, and Jacob D Durtschi. Next-generation sequencing: From basic research to diagnostics. *Clinical Chemistry*, 55(4):641–658, 2009.

[79] Shanika L Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21(1):30, 2020.

[80] Glennis A Logsdon, Mitchell R Vollger, and Evan E Eichler. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10):597–614, 2020.

[81] Mohit K Midha, Mengchu Wu, and Kuo-Ping Chiu. Long-read sequencing in deciphering human genetics to a greater depth. *Human Genetics*, 138(11-12): 1201–1215, 2019.

[82] Michael Eisenstein. Startups use short-read data to expand long-read sequencing market. *Nature Biotechnology*, 33(5):433–436, 2015.

[83] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

[84] Benilton S Carvalho and Rafael A Irizarry. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19):2363–2367, 2010.

[85] Laurent Gautier, Leslie Cope, Benjamin M Bolstad, and Rafael A Irizarry. Affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20 (3):307–315, 2004.

[86] John Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427, 2001.

[87] Dries Knapen, Lucia Vergauwen, Kris Laukens, and Ronny Blust. Best practices for hybridization design in two-colour microarray analysis. *Trends in Biotechnology*, 27(7):406–414, 2009.

[88] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[89] Leming Shi, Wendell Jones, Roderick Jensen, Stephen Harris, Roger Perkins, Federico Goodsaid, Lei Guo, Lisa Croner, Cecilie Boysen, and Hong Fang. The reproducibility of lists of differentially expressed genes in microarray studies. *Nature Precedings*, pages 1–1, 2007.

[90] G Wesley Hatfield, She-pin Hung, and Pierre Baldi. Differential analysis of dna microarray gene expression data. *Molecular Microbiology*, 47(4):871–877, 2003.

[91] Mark R Dalman, Anthony Deeter, Gayathri Nimishakavi, and Zhong-Hui Duan. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics*, 13(Suppl 2):S11, 2012.

[92] Thomas J Mariani, Vikram Budhraja, Brigham H Mecham, C Charles Gu, Mark A Watson, and Yoel Sadovsky. A variable fold-change threshold determines significance for expression microarrays. *The FASEB Journal*, 17(2):321–323, 2003.

[93] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, and Xuegong Zhang. A survey of best practices for rna-seq data analysis. *Genome biology*, 17:13, 2016.

[94] Brian J Haas and Michael C Zody. Advancing rna-seq analysis. *Nature Biotechnology*, 28(5):421–423, 2010.

[95] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[96] Steven W Wingett and Simon Andrews. Fastq screen: A tool for multi-genome mapping and quality control. *F1000Research*, 7:1338, 2018.

[97] Minh Duc Cao, Trevor I Dix, and Lloyd Allison. A genome alignment algorithm based on compression. *BMC Bioinformatics*, 11:599, 2010.

[98] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[99] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: Discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[100] Yang Liao, Gordon K Smyth, and Wei Shi. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10): e108–e108, 2013.

[101] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: Ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[102] Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nature Biotechnology*, 37(8):907–915, 2019.

[103] Thomas D Wu, Jens Reeder, Michael Lawrence, Gabe Becker, and Matthew J Brauer. Gmap and gsnap for genomic sequence alignment: Enhancements to speed, accuracy, and functionality. *Methods in Molecular Biology (Clifton, N.J.)*, 1418:283–334, 2016.

[104] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature Protocols*, 7(3):562–578, 2012.

[105] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature Biotechnology*, 33(3):290–295, 2015.

[106] Lorian Schaeffer, Harold Pimentel, Nicolas Bray, Páll Melsted, and Lior Pachter. Pseudoalignment for metagenomic read assignment. *Bioinformatics*, 33(14): 2082–2088, 2017.

[107] Fatemeh Almodaresi, Hirak Sarkar, Avi Srivastava, and Rob Patro. A space and time-efficient index for the compacted colored de bruijn graph. *Bioinformatics*, 34(13):i169–i177, 2018.

[108] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016.

[109] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, 2017.

[110] Zhide Fang and Xiangqin Cui. Design and validation issues in rna-seq experiments. *Briefings in Bioinformatics*, 12(3):280–287, 2011.

[111] Yingdong Zhao, Ming-Chung Li, Mariam M Konaté, Li Chen, Biswajit Das, Chris Karlovich, P Mickey Williams, Yvonne A Evrard, James H Doroshow, and Lisa M McShane. Tpm, fpkm, or normalized counts? a comparative study of quantification measures for the analysis of rna-seq data from the nci patient-derived models repository. *Journal of Translational Medicine*, 19:269, 2021.

[112] Zainab Khatoon, Bryan Figler, Hui Zhang, and Feng Cheng. Introduction to rna-seq and its applications to drug discovery and development. *Drug Development Research*, 75(5):324–330, 2014.

[113] David Sims, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting. Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, 2014.

[114] Shanrong Zhao, Zhan Ye, and Robert Stanton. Misuse of rpkm or tpm normalization when comparing across samples and sequencing protocols. *RNA*, 26(8): 903–909, 2020.

[115] Samuel Marguerat and Jürg Bähler. Rna-seq: From technology to biology. *Cellular and Molecular Life Sciences*, 67:569–579, 2010.

[116] Wei Zheng, Lisa M Chung, and Hongyu Zhao. Bias detection and correction in rna-sequencing data. *BMC Bioinformatics*, 12:290, 2011.

[117] Sheng Li, Paweł P Łabaj, Paul Zumbo, Peter Sykacek, Wei Shi, Leming Shi, John Phan, Po-Yen Wu, May Wang, and Charles Wang. Detecting and correcting systematic variation in large-scale rna sequencing data. *Nature Biotechnology*, 32 (9):888–895, 2014.

[118] Shanrong Zhao, Li Xi, and Baohong Zhang. Union exon based approach for rna-seq gene quantification: To be or not to be? *PloS One*, 10(11):e0141910, 2015.

[119] Matthew Dapas, Manoj Kandpal, Yingtao Bi, and Ramana V Davuluri. Comparative evaluation of isoform-level gene expression estimation algorithms for rna-seq and exon-array platforms. *Briefings in Bioinformatics*, 18(2):260–269, 2017.

[120] Charlotte Soneson, Michael I Love, and Mark D Robinson. Differential analyses for rna-seq: Transcript-level estimates improve gene-level inferences. *F1000Research*, 4:1521, 2015.

[121] Michael I Love, Simon Anders, Vladislav Kim, and Wolfgang Huber. Rna-seq workflow: Gene-level exploratory analysis and differential expression. *F1000Research*, 4:1070, 2015.

[122] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature Biotechnology*, 31(1):46–53, 2013.

[123] Chester Ittner Bliss and Ronald A Fisher. Fitting the negative binomial distribution to biological data. *Biometrics*, 9(2):176–200, 1953.

[124] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[125] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. Edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[126] Xiangqin Cui, JT Gene Hwang, Jing Qiu, Natalie J Blades, and Gary A Churchill. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6(1):59–75, 2005.

[127] Mark J Van Der Laan and Jenny Bryan. Gene expression analysis with the parametric bootstrap. *Biostatistics*, 2(4):445–461, 2001.

[128] A Stupnikov, CE McInerney, KI Savage, SA McIntosh, F Emmert-Streib, R Kennedy, M Salto-Tellez, KM Prise, and DG McArt. Robustness of differential gene expression analysis of rna-seq. *Computational and Structural Biotechnology Journal*, 19:3470–3481, 2021.

[129] Wei Pan. On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics*, 19 (11):1333–1340, 2003.

[130] Yao Wang, Chunguo Wu, Zhaohua Ji, Binghong Wang, and Yanchun Liang. Non-parametric change-point method for differential gene expression detection. *PloS One*, 6(5):e20060, 2011.

[131] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[132] Harold Pimentel, Nicolas L Bray, Suzette Puente, Páll Melsted, and Lior Pachter. Differential analysis of rna-seq incorporating quantification uncertainty. *Nature Methods*, 14(7):687–690, 2017.

[133] Sergiusz Wesolowski, Marc R Birtwistle, and Grzegorz A Rempala. A comparison of methods for rna-seq differential expression analysis and a new empirical bayes approach. *Biosensors*, 3(3):238–258, 2013.

[134] Alyssa C Frazee, Geo Pertea, Andrew E Jaffe, Ben Langmead, Steven L Salzberg, and Jeffrey T Leek. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nature Biotechnology*, 33(3):243–246, 2015.

[135] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, 2003.

[136] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome Biology*, 14:3158, 2013.

[137] Kin-Ying To. Identification of differential gene expression by high throughput analysis. *Combinatorial Chemistry & High Throughput Screening*, 3(3):235–241, 2000.

[138] Shiyi Liu, Zitao Wang, Ronghui Zhu, Feiyan Wang, Yanxiang Cheng, and Yeqiang Liu. Three differential expression analysis methods for rna sequencing: Limma, edger, deseq2. *Journal of Visualized Experiments*, 3(175):e62528, 2021.

[139] Gene Ontology Consortium. The gene ontology (go) database and informatics resource. *Nucleic Acids Research*, 32(suppl_1):D258–D261, 2004.

[140] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[141] Elise AR Serin, Harm Nijveen, Henk WM Hilhorst, and Wilco Ligterink. Learning from co-expression networks: Possibilities and challenges. *Frontiers in Plant Science*, 7:444, 2016.

[142] David Amar, Hershel Safer, and Ron Shamir. Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Computational Biology*, 9(3):e1002955, 2013.

[143] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4:Article17, 2005.

[144] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

[145] Peter Langfelder, Paul S Mischel, and Steve Horvath. When is hub gene selection better than standard meta-analysis? *PloS One*, 8(4):e61505, 2013.

[146] Wanli Yang, Jiaojiao Ma, Wei Zhou, Zichao Li, Xin Zhou, Bo Cao, Yujie Zhang, Jinqiang Liu, Zhiping Yang, and Hongwei Zhang. Identification of hub genes and outcome in colon cancer based on bioinformatics analysis. *Cancer Management and Research*, 11:323–338, 2018.

[147] Sepideh Dashti, Mohammad Taheri, and Soudeh Ghafouri-Fard. An in-silico method leads to recognition of hub genes and crucial pathways in survival of patients with breast cancer. *Scientific Reports*, 10(1):18770, 2020.

[148] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, and Michelle Holko. Ncbi geo: archive for functional genomics data sets. *Nucleic Acids Research*, 41(D1):D991–D995, 2012.

[149] Sean Davis and Paul S Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847, 2007.

[150] H Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Niran Abeygunawardena, Sergio Contrino, Richard Coulson, Anna Farne, G Garcia Lara, Ele Holloway, and Misha Kapushesky. Arrayexpress—a public repository for microarray gene expression data at the ebi. *Nucleic Acids Research*, 33(suppl_1):D553–D555, 2005.

[151] Natalie de Souza. The encode project. *Nature Methods*, 9(11):1046–1046, 2012.

[152] Eric W Sayers, Jeffrey Beck, Evan E Bolton, Devon Bourexis, James R Brister, Kathi Canese, Donald C Comeau, Kathryn Funk, Sunghwan Kim, and William Klimke. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 49(D1):D10, 2021.

[153] Rasko Leinonen, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Neil Goodgame, and Richard Gibson. The european nucleotide archive. *Nucleic Acids Research*, 39 (suppl_1):D28–D31, 2010.

[154] Jun Mashima, Yuichi Kodama, Takehide Kosuge, Takatomo Fujisawa, Toshiaki Katayama, Hideki Nagasaki, Yoshihiro Okuda, Eli Kaminuma, Osamu Ogasawara, and Kousaku Okubo. Dna data bank of japan (ddbj) progress report. *Nucleic Acids Research*, 44(D1):D51–D57, 2016.

[155] Joseph Brown, Meg Pirrung, and Lee Ann McCue. Fqc dashboard: Integrates fastqc results into a web-based, interactive, and extensible fastq quality control tool. *Bioinformatics*, 33(19):3137–3139, 2017.

[156] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. Multiqc: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016.

[157] Sylvain Pradervand, Johann Weber, Jérôme Thomas, Manuel Bueno, Pratyaksha Wirapati, Karine Lefort, G Paolo Dotto, and Keith Harshman. Impact of normalization on mirna microarray expression profiling. *RNA*, 15(3):493–501, 2009.

[158] Songbo Huang, Jinbo Zhang, Ruiqiang Li, Wenqian Zhang, Zengquan He, Tak-Wah Lam, Zhiyu Peng, and Siu-Ming Yiu. Soapsplice: Genome-wide ab initio detection of splice junctions from rna-seq data. *Frontiers in Genetics*, 2:46, 2011.

[159] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31 (2):166–169, 2015.

[160] Yang Liao, Gordon K Smyth, and Wei Shi. Featurecounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[161] Brad T Sherman, Ming Hao, Ju Qiu, Xiaoli Jiao, Michael W Baseler, H Clifford Lane, Tomozumi Imamichi, and Weizhong Chang. David: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research*, 50(W1):W216–W221, 2022.

[162] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. Clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a Journal of Integrative Biology*, 16(5):284–287, 2012.

[163] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14: 128, 2013.

[164] Guangchuang Yu and Qing-Yu He. Reactomepa: an r/bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, 12(2): 477–479, 2016.

[165] Steven Maere, Karel Heymans, and Martin Kuiper. Bingo: A cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, 2005.

[166] Jing Shi and Michael G Walker. Gene set enrichment analysis (gsea) for interpreting gene expression profiles. *Current Bioinformatics*, 2(2):133–137, 2007.

[167] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, and Peer Bork. String v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 2019.

[168] Cory D Hirsch, Nathan M Springer, and Candice N Hirsch. Genomic limitations to rna sequencing expression profiling. *The Plant Journal*, 84(3):491–503, 2015.

[169] Matthew G Hanna, Anil Parwani, and Sahussapont Joseph Sirintrapun. Whole slide imaging: Technology and applications. *Advances in Anatomic Pathology*, 27(4):251–259, 2020.

[170] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific Reports*, 6:27988, 2016.

[171] Elizabeth Munch. A user's guide to topological data analysis. *Journal of Learning Analytics*, 4(2):47–61, 2017.

[172] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6:17, 2017.

[173] Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences of the United States of America*, 108(17):7265–7270, 2011.

[174] Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532, 2018.

[175] Andrew Shapanis, Mark G Jones, James Schofield, and Paul Skipp. Topological data analysis identifies molecular phenotypes of idiopathic pulmonary fibrosis. *Thorax*, 78(7):682–689, 2023.

[176] Ruben Vandaele, Pallavi Mukherjee, Hannah M Selby, Rajan P Shah, and Olivier Gevaert. Topological data analysis of thoracic radiographic images shows improved radiomics-based lung tumor histology prediction. *Patterns*, 4(1):100657, 2023.

[177] Facundo Mémoli and Kritika Singhal. A primer on persistent homology of finite metric spaces. *Bulletin of Mathematical Biology*, 81:2074–2116, 2019.

[178] Michael Kerber. Persistent homology - state of the art and challenges. *Internationale Mathematische Nachrichten*, 70(231):15–33, 2016.

[179] Mathieu Carriere and Andrew Blumberg. Multiparameter persistence image for topological machine learning. *Advances in Neural Information Processing Systems*, 33:22432–22444, 2020.

[180] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(1):77–102, 2015.

[181] Paul Bendich, Herbert Edelsbrunner, and Michael Kerber. Computing robustness and persistence for images. *IEEE Transactions on Visualization and cComputer Graphics*, 16(6):1251–1260, 2010.

[182] Bing Liu. Supervised learning. In *Web Data Mining. Data-Centric Systems and Applications*, pages 63–132. Springer, Berlin, Heidelberg, 2011.

[183] Weixin Yao and Longhai Li. A new regression model: Modal linear regression. *Scandinavian Journal of Statistics*, 41(3):656–671, 2014.

[184] Mohamed Alloghani, Dhiya Al-Jumeily, Jamila Mustafina, Abir Hussain, and Ahmed J Aljaaf. A systematic review on supervised and unsupervised machine learning algorithms for data science. In Michael Berry, Ammar Mohamed, and Bee Wah Yap, editors, *Supervised and Unsupervised Learning for Data Science*. Springer, Cham, 2020.

[185] K Sindhu Meena and S Suriya. A survey on supervised and unsupervised learning techniques. In L. Kumar, L. Jayashree, and R. Manimegalai, editors, *Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications. AISGSC 2019 2019*, pages 627–644. Springer, Cham, 2020.

[186] Guangyu Robert Yang and Xiao-Jing Wang. Artificial neural networks for neuroscientists: A primer. *Neuron*, 107(6):1048–1070, 2020.

[187] Weipeng Cao, Xizhao Wang, Zhong Ming, and Jinzhu Gao. A review on neural networks with random weights. *Neurocomputing*, 275:278–287, 2018.

[188] John T Hancock and Taghi M Khoshgoftaar. Survey on categorical data for neural networks. *Journal of Big Data*, 7:28, 2020.

[189] Mohaiminul Islam, Guorong Chen, and Shangzhu Jin. An overview of neural network. *American Journal of Neural Networks and Applications*, 5(1):7–11, 2019.

[190] Charlie Tang and Russ R Salakhutdinov. Learning stochastic feedforward neural networks. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'13)*, pages 530–538, Red Hook, NY, USA, 2013. Curran Associates Inc.

[191] Jörg Behler. Four generations of high-dimensional neural network potentials. *Chemical Reviews*, 121(16):10037–10072, 2021.

[192] Anamika Dhillon and Gyanendra K Verma. Convolutional neural network: A review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence*, 9(2):85–112, 2020.

[193] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, and Jianfei Cai. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

[194] Jianxin Wu. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, 5(23):495, 2017.

[195] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):6999–7019, 2022.

[196] Teresa Araújo, Guilherme Aresta, Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia, and Aurélio Campilho. Classification of breast cancer histology images using convolutional neural networks. *PloS One*, 12(6): e0177544, 2017.

[197] Dhanesh Ramachandram and Graham W Taylor. A deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34 (6):96–108, 2017.

[198] Abbas Pourhossein Gilakjani, Hairul Nizam Ismail, and Seyedeh Masoumeh Ahmadi. The effect of multimodal learning models on language teaching and learning. *Theory & Practice in Language Studies*, 1(10), 2011.

[199] Francisco Carrillo-Perez, Juan Carlos Morales, Daniel Castillo-Secilla, Yésica Molina-Castro, Alberto Guillén, Ignacio Rojas, and Luis Javier Herrera. Non-small-cell lung cancer classification via rna-seq and histology imaging probability fusion. *BMC Bioinformatics*, 22(1):454, 2021.

[200] Karl Y Bilimoria, Andrew K Stewart, David P Winchester, and Clifford Y Ko. The national cancer data base: A powerful initiative to improve cancer care in the united states. *Annals of Surgical Oncology*, 15:683–690, 2008.

[201] Melissa S Cline, Brian Craft, Teresa Swatloski, Mary Goldman, Singer Ma, David Haussler, and Jingchun Zhu. Exploring tcga pan-cancer data at the ucsc cancer genomics browser. *Scientific Reports*, 3(1):2652, 2013.

[202] Barbara A Conley, Lou Staudt, Naoko Takebe, David A Wheeler, Linghua Wang, Maria F Cardenas, Viktoriya Korchina, Jean Claude Zenklusen, Lisa M McShane, and James V Tricoli. The exceptional responders initiative: Feasibility of a national cancer institute pilot study. *JNCI: Journal of the National Cancer Institute*, 113(1):27–37, 2021.

[203] Go Nishikawa, Jia Luo, and Vinay Prasad. A comprehensive review of exceptional responders to anticancer drugs in the biomedical literature. *European Journal of Cancer*, 101:143–151, 2018.

[204] Oluwadamilola Olaku, Barbara A Conley, S Percy Ivy, Lisa M McShane, Louis M Staudt, Sophie M King, Megan Sansevere, Benjamin Kim, and Jeffrey D White. Survey of lifestyle, past medical history and complementary and alternative medicine use among adult patients participating in the national cancer institute's exceptional responders initiative. *Translational Oncology*, 25:101484, 2022.

[205] Anders B Dohlman, Jared Klug, Marissa Mesko, Iris H Gao, Steven M Lipkin, Xiling Shen, and Iliyan D Iliev. A pan-cancer mycobiome analysis reveals fungal involvement in gastrointestinal and lung tumors. *Cell*, 185(20):3807–3822, 2022.

[206] Allison Warren, Yejia Chen, Andrew Jones, Tsukasa Shibue, William C Hahn, Jesse S Boehm, Francisca Vazquez, Aviad Tsherniak, and James M McFarland. Global computational alignment of tumor and cell line transcriptional profiles. *Nature Communications*, 12(1):22, 2021.

[207] Jiannan Liu, Chuanpeng Dong, Guanglong Jiang, Xiaoyu Lu, Yunlong Liu, and Huanmei Wu. Transcription factor expression as a predictor of colon cancer prognosis: A machine learning practice. *BMC Medical Genomics*, 13(Suppl 9): 135, 2020.

[208] Dejun Wu, Zhenhua Yin, Yisheng Ji, Lin Li, Yunxin Li, Fanqiang Meng, Xiaohan Ren, and Ming Xu. Identification of novel autophagy-related lncrnas associated with a poor prognosis of colon adenocarcinoma through bioinformatics analysis. *Scientific Reports*, 11(1):8069, 2021.

[209] Ehsan Akbari Sekehravani, Eduard Babulak, and Mehdi Masoodi. Implementing canny edge detection algorithm for noisy image. *Bulletin of Electrical Engineering and Informatics*, 9(4):1404–1410, 2020.

[210] Baştan, Muhammet and Bukhari, Syed Saqib and Breuel, Thomas. Active canny: Edge detection and recovery with open active contour models. *IET Image Processing*, 11(12):1325–1332, 2017.

[211] Ahmed Shihab Ahmed. Comparative study among sobel, prewitt and canny edge detection operators used in image processing. *Journal of Theoretical and Applied Information Technology*, 96(19):6517–6525, 2018.

[212] MH Shokhan. An efficient approach for improving canny edge detection algorithm. *International Journal of Advances in Engineering & Technology*, 7(1):59, 2014.

[213] Renjie Song, Ziqi Zhang, and Haiyang Liu. Edge connection based canny edge detection algorithm. *Pattern Recognition and Image Analysis*, 27:740–747, 2017.

[214] Ping Zhou, Wenjun Ye, Yaojie Xia, and Qi Wang. An improved canny algorithm for edge detection. *Journal of Computational Information Systems*, 7(5):1516–1523, 2011.

[215] Dan Dan Han, Tian Chi Zhang, and Jing Zhang. Research and implementation of an improved canny edge detection algorithm. *Key Engineering Materials*, 572: 566–569, 2014.

[216] Romanus Damanik. Image detection edge image using canny edge algorithm. *Journal of Artificial Intelligence and Engineering Applications*, 1(3):248–255, 2022.

[217] Ranjan Muthukrishnan and Miyilsamy Radha. Edge detection techniques for image segmentation. *International Journal of Computer Science & Information Technology*, 3(6):259, 2011.

[218] Manuel Muñoz-Aguirre, Vasilis F Ntasis, Santiago Rojas, and Roderic Guigó. Pyhist: A histological image segmentation tool. *PLoS Computational Biology*, 16 (10):e1008349, 2020.

[219] Marijo Bilusic, Daniel Girardi, Yan Zhou, Kyungsuk Jung, Jianming Pei, Michael Slifker, Qingrong Chen, Daoud Meerzaman, Katherine Alpaugh, and Denise Young. Molecular profiling of exceptional responders to cancer therapy. *The Oncologist*, 26(3):186–195, 2021.

[220] Ergang Wang, Kun Xiang, Yun Zhang, and Xiao-Fan Wang. Patient derived organoids (pdos) and pdo-derived xenografts (pdoxs): New opportunities in establishing faithful preclinical cancer models. *Journal of the National Cancer Center*, 2022.

[221] SM Lin, XJ Wang, SH Huang, ZB Xu, Y Huang, XR Lu, DB Xu, and P Chi. Construction of artificial neural network model for predicting the efficacy of first-line folfox chemotherapy for metastatic colorectal cancer. *Chinese Journal of Oncology*, 43(2):202–206, 2021.

[222] Lei Zhang, Chao Xu, Si-Han Wang, Qin-Wen Ge, Xiao-Wei Wang, Pan Xiao, and Qing-Hua Yao. Cancer-associated fibroblast-related gene signatures predict survival and drug response in patients with colorectal cancer. *Frontiers in Genetics*, 13:1054152, 2022.

[223] Xiao-Shun He, Sheng-Yi Zou, Jia-Lu Yao, Wangjianfei Yu, Zhi-Yong Deng, Jing-Ru Wang, Wen-Juan Gan, Shan Wan, Xiao-Qin Yang, and Hua Wu. Transcriptomic analysis identifies complement component 3 as a potential predictive biomarker for chemotherapy resistance in colorectal cancer. *Frontiers in Molecular Biosciences*, 8:763652, 2021.

[224] K McHugh and S Kao. Response evaluation criteria in solid tumours (recist): Problems and need for modifications in paediatric oncology? *The British Journal of Radiology*, 76(907):433–436, 2003.

[225] Yimin Li, Yaqi Gan, Jiaxin Liu, Juanni Li, Zhengwei Zhou, Ruotong Tian, Ruizheng Sun, Jiaqi Liu, Qing Xiao, and Yuanyuan Li. Downregulation of meis1 mediated by elfn1-as1/ezh2/dnmt3a axis promotes tumorigenesis and oxaliplatin resistance in colorectal cancer. *Signal Transduction and Targeted Therapy*, 7(1): 87, 2022.

[226] Lingyu Han, Weixing Dai, Wenqin Luo, Li Ye, Hongsheng Fang, Shaobo Mo, Qingguo Li, Ye Xu, Renjie Wang, and Guoxiang Cai. Enhanced de novo lipid synthesis mediated by fasn induces chemoresistance in colorectal cancer. *Cancers*, 15(3):562, 2023.

[227] Christina Demetriadou, Anastasia Raoukka, Evelina Charidemou, Constantine Mylonas, Christina Michael, Swati Parekh, Costas Koufaris, Paris Skourides, Panagiotis Papageorgis, and Peter Tessarz. Histone n-terminal acetyltransferase naa40 links one-carbon metabolism to chemoresistance. *Oncogene*, 41(4):571–585, 2022.

[228] Hufei Wang, Zhi Li, Suwen Ou, Yanni Song, Kangjia Luo, Zilong Guan, Lei Zhao, Rui Huang, and Shan Yu. Tumor microenvironment heterogeneity-based score system predicts clinical prognosis and response to immune checkpoint blockade in

multiple colorectal cancer cohorts. *Frontiers in Molecular Biosciences*, 9:884839, 2022.

[229] Cheng-Zhi Huang, Yue Zhou, Qi-Song Tong, Qi-Jia Duan, Qing Zhang, Jin-Zhi Du, and Xue-Qing Yao. Precision medicine-guided co-delivery of aspn sirna and oxaliplatin by nanoparticles to overcome chemoresistance of colorectal cancer. *Biomaterials*, 290:121827, 2022.

[230] Yafei Zhang, Jianqiong Lin, Kaibin Yang, and Zhicao Yue. Chemotherapy suppresses shh gene expression via a specific enhancer. *Journal of Genetics and Genomics*, 50(1):27–37, 2023.

[231] Hamed Manoochehri, Hamid Tanzadehpanah, Amir Taherkhani, and Akram Jalali. A gene expression signature to predict chemotherapy response of colorectal cancer patients: Systems biology analysis on transcriptomics data and experimental validation. *Journal of Applied Biotechnology Reports*, 9(2):657–670, 2022.

[232] José A Díaz-García and Graciela González-Farías. A note on cook's distance. *Journal of Statistical Planning and Inference*, 120(1-2):119–136, 2004.

[233] Myung Geun Kim. A cautionary note on the use of cook's distance. *Communications for Statistical Applications and Methods*, 24(3):317–324, 2017.

[234] GS David Sam Jayakumar and A Sulthan. Exact distribution of cook's distance and identification of influential observations. *Hacettepe Journal of Mathematics and Statistics*, 44(1):165–178, 2015.

[235] Danni Yu, Wolfgang Huber, and Olga Vitek. Shrinkage estimation of dispersion in negative binomial models for rna-seq experiments with small sample size. *Bioinformatics*, 29(10):1275–1282, 2013.

[236] Clarissa M Koch, Stephen F Chiu, Mahzad Akbarpour, Ankit Bharat, Karen M Ridge, Elizabeth T Bartom, and Deborah R Winter. A beginner's guide to analysis of rna sequencing data. *American Journal of Respiratory Cell and Molecular Biology*, 59(2):145–157, 2018.

[237] Nicholas J Schurch, Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G Simpson, and Tom Owen-Hughes. How many biological replicates are needed in an rna-seq experiment and which differential expression tool should you use? *RNA*, 22(6): 839–851, 2016.

[238] Wolfgang Huber, Anja Von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(suppl_1): S96–S104, 2002.

[239] Bronwen L Aken, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos García Girón, and Thibaut Hourlier. The ensembl gene annotation system. *Database*, 2016: baw093, 2016.

[240] Allison Piovesan, Francesca Antonaros, Lorenza Vitale, Pierluigi Strippoli, Maria Chiara Pelleri, and Maria Caracausi. Human protein-coding genes and gene feature statistics in 2019. *BMC Research Notes*, 12:315, 2019.

[241] JT Gene Hwang, Jing Qiu, and Zhigen Zhao. Empirical bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(1):265–285, 2009.

[242] David Gerard and Matthew Stephens. Empirical bayes shrinkage and false discovery rate estimation, allowing for unwanted variation. *Biostatistics*, 21(1): 15–32, 2020.

[243] Debashis Ghosh. Incorporating the empirical null hypothesis into the benjamini-hochberg procedure. *Statistical Applications in Genetics and Molecular Biology*, 11(4), 2012.

[244] Miri Adler and Uri Alon. Fold-change detection in biological systems. *Current Opinion in Systems Biology*, 8:81–89, 2018.

[245] Gwowen Shieh. On power and sample size calculations for wald tests in generalized linear models. *Journal of Statistical Planning and Inference*, 128(1):43–59, 2005.

[246] Jianhua Ruan, Angela K Dean, and Weixiong Zhang. A general co-expression network-based approach to gene expression analysis: Comparison and applications. *BMC Systems Biology*, 4:8, 2010.

[247] Daniele Mercatelli, Laura Scalambra, Luca Triboli, Forest Ray, and Federico M Giorgi. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6): 194430, 2020.

[248] Peter Langfelder and Steve Horvath. Fast r functions for robust correlations and hierarchical clustering. *Journal of Statistical Software*, 46(11), 2012.

[249] Patrick Veenstra, Colin Cooper, and Steve Phelps. The use of biweight mid correlation to improve graph based portfolio construction. In *2016 8th Computer Science and Electronic Engineering (CEEC)*, pages 101–106, Colchester, UK, 2016.

[250] Mark EJ Newman. Analysis of weighted networks. *Physical Review E*, 70(5): 056131, 2004.

[251] Andy M Yip and Steve Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, 8:22, 2007.

[252] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[253] Yingling Wan, Min Zhang, Aiying Hong, Xinyu Lan, Huiyan Yang, and Yan Liu. Transcriptome and weighted correlation network analyses provide insights into inflorescence stem straightness in paeonia lactiflora. *Plant Molecular Biology*, 102:239–252, 2020.

[254] Xia Chen and Jianfeng Ma. Weighted gene co-expression network analysis (wgcna) to explore genes responsive to streptococcus oralis biofilm and immune infiltration analysis in human gingival fibroblasts cells. *Bioengineered*, 12(1): 1054–1065, 2021.

[255] Y. Wei, Q. Su, and X. Li. Identification of hub genes related to duchenne muscular dystrophy by weighted gene co-expression network analysis. *Medicine (Baltimore)*, 101(52):e32603, 2022.

[256] Qisheng Su, Qinpei Ding, Zunni Zhang, Zheng Yang, Yuling Qiu, Xiaohong Li, and Wuning Mo. Identification of genes associated with the metastasis of pheochromocytoma/paraganglioma based on weighted gene coexpression network analysis. *BioMed Research International*, 2020:3876834, 2020.

[257] Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for r. *Bioinformatics*, 24 (5):719–720, 2008.

[258] Zhixin Wu, Yinxian Wen, Guanlan Fan, Hangyuan He, Siqi Zhou, and Liaobin Chen. Hemgn and slc2a1 might be potential diagnostic biomarkers of steroid-induced osteonecrosis of femoral head: study based on wgcna and degs screening. *BMC Musculoskeletal Disorders*, 22(85):85, 2021.

[259] Kun-Hao Bai, Si-Yuan He, Ling-Ling Shu, Wei-Da Wang, Shi-Yong Lin, Qian-Yi Zhang, Liang Li, Lei Cheng, and Yu-Jun Dai. Identification of cancer stem cell characteristics in liver hepatocellular carcinoma by wgcna analysis of transcriptome stemness index. *Cancer Medicine*, 9(12):4290–4298, 2020.

[260] Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Research*, 32:D258–D261, 2004.

[261] Charles A Tilford and Nathan O Siemers. Gene set enrichment analysis. *Methods in molecular biology (Clifton, N.J.)*, 563:99–121, 2009.

[262] Wenting Liu, Jianjun Liu, and Jagath C Rajapakse. Gene ontology enrichment improves performances of functional similarity of genes. *Scientific Reports*, 8(1): 12100, 2018.

[263] Tao Zhou, Jun Yao, and Zhanjiang Liu. Gene ontology, enrichment analysis, and pathway analysis. *Bioinformatics in Aquaculture: Principles and Methods*, pages 150–168, 2017.

[264] Quang Hung Nguyen, Hai-Bang Ly, Lanh Si Ho, Nadhir Al-Ansari, Hiep Van Le, Van Quan Tran, Indra Prakash, and Binh Thai Pham. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, 2021:Article ID 4832864, 2021.

[265] Vikash Singh, Michael Pencina, Andrew J Einstein, Joanna X Liang, Daniel S Berman, and Piotr Slomka. Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Scientific Reports*, 11(1):14490, 2021.

[266] Qing-Song Xu and Yi-Zeng Liang. Monte carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1):1–11, 2001.

[267] Ekin Yagis, Selamawet Workalemahu Atnafu, Alba García Seco de Herrera, Chiara Marzi, Riccardo Scheda, Marco Giannelli, Carlo Tessa, Luca Citi, and Stefano Diciotti. Effect of data leakage in brain mri classification using 2d convolutional neural networks. *Scientific Reports*, 11(1):22544, 2021.

[268] Philip Sedgwick. Pearson's correlation coefficient. *BMJ*, 345, 2012.

[269] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

[270] Carlos Ramos-Carreño and José L Torrecilla. Dcor: Distance correlation and energy statistics in python. *SoftwareX*, 22:101326, 2023.

[271] Xueqin Wang, Wenliang Pan, Wenhao Hu, Yuan Tian, and Heping Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015.

[272] Emily Greenfest-Allen, Jean-Philippe Cartailler, Mark A. Magnuson, and Christian J. Stoeckert. iterativewgcna: iterative refinement to improve module detection from wgcna co-expression networks. *bioRxiv*, 2017. URL https://api.semanticscholar.org/CorpusID:89699883.

[273] Mehmet E Aktas, Esra Akbas, and Ahmed El Fatmaoui. Persistence homology of networks: Methods and applications. *Applied Network Science*, 4(1):1–28, 2019.

[274] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(1):218–252, 2017. ISSN 1532-4435.

[275] C. Maria, JD. Boissonnat, M. Glisse, and M. Yvinec. The gudhi library: Simplicial complexes and persistent homology. In H. Hong and C. Yap, editors, *Mathematical Software – ICMS 2014*, volume 8592 of *Lecture Notes in Computer Science*, Berlin, Heidelberg, 2014. Springer.

[276] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861, 2017. URL https://api.semanticscholar.org/CorpusID:12670695.

[277] Budhaditya Saha, Sunil Gupta, Dinh Phung, and Svetha Venkatesh. Multiple task transfer learning with small sample sizes. *Knowledge and Information Systems*, 46:315–342, 2016.

[278] Samir Al-Stouhi and Chandan K Reddy. Transfer learning for class imbalance problems with inadequate data. *Knowledge and Information Systems*, 48:201–228, 2016.

[279] Junde Chen, Defu Zhang, and Yaser Ahangari Nanehkaran. Identifying plant diseases using deep transfer learning and enhanced lightweight network. *Multimedia Tools and Applications*, 79:31497–31515, 2020.

[280] Nikhil Ketkar and Nikhil Ketkar. Introduction to keras. In *Deep Learning with Python*, pages 97–111. Apress, Berkeley, CA, 2017.

[281] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, and Matthieu Devin. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *Computing Research Repository*, ArXiv:1603.04467, 2016.

[282] Hercules Dalianis and Hercules Dalianis. Evaluation metrics and evaluation. In *Clinical Text Mining*, pages 45–53. Springer, Cham, 2018.

[283] Kanae Takahashi, Kouji Yamamoto, Aya Kuchiba, and Tatsuki Koyama. Confidence interval for micro-averaged f1 and macro-averaged f1 scores. *Applied Intelligence*, 52(5):4961–4972, 2022.

[284] Kun Zeng. Automated classification of clinical trial eligibility criteria text based on ensemble learning and metric learning. *BMC Medical Informatics and Decision Making*, 21(Suppl 2):129, 2021.

[285] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.

[286] Michael Buckland and Fredric Gey. The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1):12–19, 1994.

[287] David Thissen. Reliability and measurement precision. In *Computerized Adaptive Testing*, pages 159–184. Routledge, 2000.

[288] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3): 299–310, 2005.

[289] Alberto Jiménez-Valverde. Insights into the area under the receiver operating characteristic curve (auc) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, 21(4):498–507, 2012.

[290] Hind Taud and Jean-François Mas. Multilayer perceptron (mlp). pages 451–455, 2018. URL https://api.semanticscholar.org/CorpusID:67464997.

[291] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[292] Michalis K. Titsias. One-vs-each approximation to softmax for scalable estimation of probabilities, 2016.

[293] Peter Flach and Meelis Kull. Precision-recall-gain curves: Pr analysis done right. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, volume 1, pages 838–846. Massachusetts Institute of Technology (MIT) Press, 2015. URL https://papers.nips.cc/paper/5867-precision-recall-gain-curves-pr-analysis-done-right.

[294] Cornelis JA Punt, Miriam Koopman, and Louis Vermeulen. From tumour heterogeneity to advances in precision treatment of colorectal cancer. *Nature Reviews Clinical Oncology*, 14(4):235–246, 2017.

[295] Said Yacine Boulahia, Abdenour Amamra, Mohamed Ridha Madi, and Said Daikh. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6): 121, 2021.

[296] Yagya Raj Pandeya and Joonwhoan Lee. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 80:2887–2905, 2021.

[297] Matthew E Ritchie, Jeremy Silver, Alicia Oshlack, Melissa Holmes, Dileepa Diyagama, Andrew Holloway, and Gordon K Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23(20):2700–2707, 2007.

[298] Mark A Jensen, Vincent Ferretti, Robert L Grossman, and Louis M Staudt. The nci genomic data commons as an engine for precision medicine. *Blood, The Journal of the American Society of Hematology*, 130(4):453–459, 2017.

[299] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

[300] Naghmeh Poorinmohammad, Jing Fu, Bob Wabeke, and Eduard J Kerkhoven. Validated growth rate-dependent regulation of lipid metabolism in yarrowia lipolytica. *International Journal of Molecular Sciences*, 23(15):8517, 2022.

[301] Miao Wang, Hongyan Gao, Weihong Li, and Bin Wu. Icariin and its metabolites regulate lipid metabolism: From effects to molecular mechanisms. *Biomedicine & Pharmacotherapy*, 131:110675, 2020.

[302] Ke Li, Ling Li, Mengliu Yang, Hua Liu, Guenther Boden, and Gangyi Yang. The effects of fibroblast growth factor-21 knockdown and over-expression on its signaling pathway and glucose–lipid metabolism in vitro. *Molecular and Cellular Endocrinology*, 348(1):21–26, 2012.

[303] Chinmayee Sethy and Chanakya Nath Kundu. 5-fluorouracil (5-fu) resistance and the new strategy to enhance the sensitivity against cancer: Implication of dna repair inhibition. *Biomedicine & Pharmacotherapy*, 137:111285, 2021.

[304] Manojkumar Bupathi and Christina Wu. Biomarkers for immune therapy in colorectal cancer: Mismatch-repair deficiency. *Journal of Gastrointestinal Oncology*, 7(5):713, 2016.

[305] Shamin Azwar, Heng Fong Seow, Maha Abdullah, Mohd Faisal Jabar, and Norhafizah Mohtarrudin. Recent updates on mechanisms of resistance to 5-fluorouracil and reversal strategies in colon cancer treatment. *Biology*, 10(9):854, 2021.

[306] Seung Tae Kim, Jeeyun Lee, Se Hoon Park, Joon Oh Park, Ho Yeong Lim, Won Ki Kang, Jin Yong Kim, Young Ho Kim, Dong Kyung Chang, and Poong-Lyul Rhee. The effect of dna mismatch repair (mmr) status on oxaliplatin-based first-line chemotherapy as in recurrent or metastatic colon cancer. *Medical Oncology*, 27: 1277–1285, 2010.

[307] Sun Tian, Fulong Wang, Shixun Lu, and Gong Chen. Identification of two subgroups of folfox. *Cancer Investigation*, 39(1):62–72, 2021.

[308] Réka Albert. Scale-free networks in cell biology. *Journal of Cell Science*, 118 (21):4947–4957, 2005.

[309] Raya Khanin and Ernst Wit. How scale-free are biological networks. *Journal of Computational Biology*, 13(3):810–818, 2006.

[310] Li M Chen, Zhixun Su, Bo Jiang, and Li M Chen. *Topological data analysis*. Springer Publishing Company, Incorporated, 1st edition, 2015. ISBN 3319251252.

[311] Jing Wu, Yifan Li, Yu Huang, Lanxiang Liu, Hanping Zhang, Corina Nagy, Xunmin Tan, Ke Cheng, Yiyun Liu, and Juncai Pu. Integrating spatial and single-nucleus transcriptomic data elucidates microglial-specific responses in female cynomolgus macaques. *Nature Neuroscience*, 26(8):1352–1364, 2023.

[312] Samuel Morabito, Fairlie Reese, Negin Rahimzadeh, Emily Miyoshi, and Vivek Swarup. hdwgcna identifies co-expression networks in high-dimensional transcriptomics data. *Cell Reports Methods*, 3(6):100498, 2023.

[313] Eugenia Radulescu, Andrew E Jaffe, Richard E Straub, Qiang Chen, Joo Heon Shin, Thomas M Hyde, Joel E Kleinman, and Daniel R Weinberger. Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain. *Molecular Psychiatry*, 25(4):791–804, 2020.

[314] Aaron Gordon, Se-Jin Yoon, Stephen S Tran, Christopher D Makinson, Jin Young Park, Jimena Andersen, Alfredo M Valencia, Steve Horvath, Xinshu Xiao, and John Rs Huguenard. Long-term maturation of human cortical organoids matches key early postnatal transitions. *Nature Neuroscience*, 24(3):331–342, 2021.

[315] Davide Degli Esposti, Christine Almunia, Marc-Antoine Guery, Natacha Koenig, Jean Armengaud, Arnaud Chaumot, and Olivier Geffard. Co-expression network analysis identifies gonad-and embryo-associated protein modules in the sentinel species gammarus fossarum. *Scientific Reports*, 9(1):7862, 2019.

[316] Ryu Hayakawa. Quantum algorithm for persistent betti numbers and topological data analysis. *Quantum*, 6:873, 2022.

[317] Luís Duarte. Betti numbers under small perturbations. *Journal of Algebra*, 594: 138–153, 2022.

[318] Benjamin Küster and Tobias Weich. Pollicott-ruelle resonant states and betti numbers. *Communications in Mathematical Physics*, 378(2):917–941, 2020.

[319] Kongjing Gu, Liang Yan, Xiang Li, Xiaojun Duan, and Jingjie Liang. Change point detection in multi-agent systems based on higher-order features. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(11):111102, 2022.

[320] Peter D Sozou, Timothy F Cootes, Christopher J Taylor, EC Di Mauro, and Andreas Lanitis. Non-linear point distribution modelling using a multi-layer perceptron. *Image and Vision Computing*, 15(6):457–463, 1997.

[321] Qi Zhao, Ze Ye, Chao Chen, and Yusu Wang. Persistence enhanced graph neural network. In *International Conference on Artificial Intelligence and Statistics*, pages 2896–2906. PMLR, 2020.

[322] Xue Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168:022022, 2019.

[323] Malgorzata Wamil, Abdelaali Hassaine, Shishir Rao, Yikuan Li, Mohammad Mamouei, Dexter Canoy, Milad Nazarzadeh, Zeinab Bidel, Emma Copland, and Kazem Rahimi. Stratification of diabetes in the context of comorbidities, using representation learning and topological data analysis. *Scientific Reports*, 13(1): 11478, 2023.

[324] Sourav Mandal, Añamaria Guzmán-Sáenz, Niina Haiminen, Sanchita Basu, and Laxmi Parida. A topological data analysis approach on predicting phenotypes from gene expression data. In Carlos Martín-Vide, Miguel Vega-Rodríguez, and Travis Wheeler, editors, *Algorithms for Computational Biology. AlCoB 2020*, volume 12099. Springer, Cham, 2020.

[325] Yu-Chiao Chiu, Hsin-I. Chen, Tianyu Zhang, Tuo Zhang, Shuyu Zhang, Aparna Gorthi, Li-San Wang, Tim H. M. Huang, and Jake Y. Chen. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med Genomics*, 12(Suppl 1):18, 2019.

[326] Pei-Chen Tsai, Tsung-Hua Lee, Kun-Chi Kuo, Fang-Yi Su, Tsung-Lu Michael Lee, Eliana Marostica, Tomotaka Ugai, Melissa Zhao, Mai Chan Lau, and Juha P Väyrynen. Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. *Nature Communications*, 14(1):2102, 2023.

[327] Hyun-Jong Jang, Ahwon Lee, J Kang, In Hye Song, and Sung Hak Lee. Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. *World Journal of Gastroenterology*, 26(40):6207, 2020.

[328] Ziyang Xu, Abhinit Verma, Usman Naveed, Samuel F Bakhoum, Pooya Khosravi, and Olivier Elemento. Deep learning predicts chromosomal instability from histopathology images. *iScience*, 24(5):102394, 2021.

[329] Mingyu Chen, Bin Zhang, Win Topatana, Jiasheng Cao, Hepan Zhu, Sarun Juengpanich, Qijiang Mao, Hong Yu, and Xiujun Cai. Classification and mutation prediction based on histopathology h&e images in liver cancer using deep learning. *NPJ Precision Oncology*, 4(1):14, 2020.

[330] Jakob Nikolas Kather, Alexander T Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H Loosen, Alexander Marx, Peter Boor, Frank Tacke, and Ulf Peter Neumann. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine*, 25(7):1054–1056, 2019.

[331] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10): 1559–1567, 2018.

[332] Jakob Nikolas Kather, Lara R Heij, Heike I Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M Niehues, Kai AJ Sommer, and Peter Bankhead. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, 1(8):789–799, 2020.

[333] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*, 1(8):800–810, 2020.

[334] Lingfeng He, Hong Zhu, Shiying Zhou, Ting Wu, Huan Wu, Huan Yang, Huiwen Mao, Chandra SekharKathera, Avilala Janardhan, and Ashlin M Edick. Wnt pathway is involved in 5-fu drug resistance of colorectal cancer cells. *Molecular Cancer*, 21(1):144, 2022.

[335] Caitriona Holohan, Sandra Van Schaeybroeck, Daniel B Longley, and Patrick G Johnston. Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer*, 13(10):714–726, 2013.

[336] Elaheh Alizadeh, Jordan Castle, Analia Quirk, Cameron DL Taylor, Wenlong Xu, and Ashok Prasad. Cellular morphological features are predictive markers of cancer cell state. *Computers in Biology and Medicine*, 126:104044, 2020.

[337] Michelle J Ziperstein, Asja Guzman, and Laura J Kaufman. Cell morphology best predicts tumorigenicity and metastasis in vivo across multiple tnbc cell lines of different metastatic potential. *Breast Cancer Research*, 26(1):43, 2024.

[338] Xuan Wang, Haiyun Zhang, and Xiaozhuo Chen. Drug resistance and combating drug resistance in cancer. *Cancer Drug Resistance*, 2(2):141, 2019.

[339] Ian Ganly, Tihana Ibrahimpasic, Michael Rivera, Ian Nixon, Frank Palmer, Snehal G Patel, R Michael Tuttle, Jatin P Shah, and Ronald Ghossein. Prognostic implications of papillary thyroid carcinoma with tall-cell features. *Thyroid*, 24 (4):662–670, 2014.

[340] John F Thompson, Seng-Jaw Soong, Charles M Balch, Jeffrey E Gershenwald, Shouluan Ding, Daniel G Coit, Keith T Flaherty, Phyllis A Gimotty, Timothy Johnson, and Marcella M Johnson. Prognostic significance of mitotic rate in localized primary cutaneous melanoma: An analysis of patients in the multi-institutional american joint committee on cancer melanoma staging database. *Journal of Clinical Oncology*, 29(16):2199, 2011.

[341] Alessandra Marconi, Marika Quadri, Francesca Farnetani, Silvana Ciardo, Elisabetta Palazzo, Roberta Lotti, Anna Maria Cesinaro, Luca Fabbiani, Cristina Vaschieri, and Mario Puviani. In vivo melanoma cell morphology reflects molecular signature and tumor aggressiveness. *Journal of Investigative Dermatology*, 142(8):2205–2216, 2022.

[342] Xuebing Wang, Zineng Xu, Yanhang Tong, Long Xia, Bimeng Jie, Peng Ding, Hailong Bai, Yi Zhang, and Yang He. Detection and classification of mandibular fracture on ct scan using deep convolutional neural network. *Clinical Oral Investigations*, 26(6):4593–4601, 2022.

[343] Jianing Zhang, Xuechen Li, Yuexiang Li, Mingyu Wang, Bingsheng Huang, Shuqiao Yao, and Linlin Shen. Three dimensional convolutional neural network-based classification of conduct disorder with structural mri. *Brain Imaging and Behavior*, 14:2333–2340, 2020.

[344] Deogratias Mzurikwao, Muhammad Usman Khan, Oluwarotimi Williams Samuel, Jindrich Cinatl Jr, Mark Wass, Martin Michaelis, Gianluca Marcelli, and Chee Siang Ang. Towards image-based cancer cell lines authentication using deep neural networks. *Scientific Reports*, 10(1):19857, 2020.

[345] Abdelbaki Souid, Nizar Sakli, and Hedi Sakli. Classification and predictions of lung diseases from chest x-rays using mobilenet v2. *Applied Sciences*, 11(6):2751, 2021.

[346] Elhoucine Elfatimi, Recep Eryigit, and Lahcen Elfatimi. Beans leaf diseases classification using mobilenet models. *IEEE Access*, 10:9471–9482, 2022.

[347] T. H. Arfan, M. Hayaty, and A. Hadinegoro. Classification of brain tumours types based on mri images using mobilenet. In *2021 2nd International Conference on Innovative and Creative Information Technology (ICITech)*, pages 69–73, Salatiga, Indonesia, 2021. doi: 10.1109/ICITech50181.2021.9590183.

[348] Shambhavi Kumar, Arohi Narang, Monica Parihar, and Vidya Sawant. Breast tumour classification using mobilenet v2. In *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pages 1–6. IEEE, 2021.

[349] HS Laxmisagar and MC Hanumantharaju. Detection of breast cancer with lightweight deep neural networks for histology image classification. *Critical Reviews in Biomedical Engineering*, 50, 2022.

[350] Vikash Sharma, Siddhartha Roy, and Girdhar G Agarwal. Breast cancer in histopathological data through image classification using deep learning methods. *Machine Learning*, 2(1):1–37, 2023.

[351] Mika S Jain and Tarik F Massoud. Predicting tumour mutational burden from histopathological images using multiscale deep learning. *Nature Machine Intelligence*, 2(6):356–362, 2020.

[352] Richard A Ward, Stephen Fawell, Nicolas Floc'h, Vikki Flemington, Darren McKerrecher, and Paul D Smith. Challenges and opportunities in cancer drug resistance. *Chemical Reviews*, 121(6):3297–3351, 2020.

[353] Nagasundaram Nagarajan, Edward KY Yapp, Nguyen Quoc Khanh Le, Balu Kamaraj, Abeer Mohammed Al-Subaie, and Hui-Yuan Yeh. Application of computational biology and artificial intelligence technologies in cancer precision drug discovery. *BioMed research international*, 2019:8427042, 2019.

[354] Jun Cheng, Jie Zhang, Yatong Han, Xusheng Wang, Xiufen Ye, Yuebo Meng, Anil Parwani, Zhi Han, Qianjin Feng, and Kun Huang. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. *Cancer Research*, 77(21):e91–e100, 2017.

[355] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115(13):E2970–E2979, 2018.

[356] Zhucheng Zhan, Zheng Jing, Bing He, Noshad Hosseini, Maria Westerhoff, Eun-Young Choi, and Lana X Garmire. Two-stage cox-nnet: biologically interpretable neural-network model for prognosis prediction and its application in liver cancer survival using histopathology and transcriptomic data. *NAR Genomics and Bioinformatics*, 3(1):lqab015, 2021.

[357] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, and Bumjin Joo. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022.

[358] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2020.

[359] Abir A Muftah, Mohammed Aleskandarany, Sultan N Sonbul, Christopher C Nolan, Maria Diez Rodriguez, Carlos Caldas, Ian O Ellis, Andrew R Green, and Emad A Rakha. Further evidence to support bimodality of oestrogen receptor expression in breast cancer. *Histopathology*, 70(3):456–465, 2017.

[360] Aras Asaad, Dashti Ali, Taban Majeed, and Rasber Rashid. Persistent homology for breast tumor classification using mammogram scans. *Mathematics*, 10(21): 4039, 2022.

[361] Monika Ray and Weixiong Zhang. Analysis of alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks. *BMC systems biology*, 4:136, 2010.

[362] Jacob Townsend, Cassie Putman Micucci, John H Hymel, Vasileios Maroulas, and Konstantinos D Vogiatzis. Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nature Communications*, 11(1):3230, 2020.

[363] Tamal K Dey, Sayan Mandal, and Soham Mukherjee. Gene expression data classification using topology and machine learning models. *BMC bioinformatics*, 22(Suppl 10):627, 2022.

[364] Chi Seng Pun, Brandon Yung Sin Yong, and Kelin Xia. Weighted-persistent-homology-based machine learning for rna flexibility analysis. *PloS One*, 15(8): e0237747, 2020.

[365] Mathieu Carriere and Andrew Blumberg. Multiparameter persistence image for topological machine learning. *Advances in Neural Information Processing Systems*, 33:22432–22444, 2020.

[366] Zhenyu Meng, D Vijay Anand, Yunpeng Lu, Jie Wu, and Kelin Xia. Weighted persistent homology for biomolecular data analysis. *Scientific Reports*, 10(1): 2079, 2020.

[367] Xiang Liu, Xiangjun Wang, Jie Wu, and Kelin Xia. Hypergraph-based persistent cohomology (hpc) for molecular representations in drug design. *Briefings in Bioinformatics*, 22(5):bbaa411, 2021.

[368] Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep learning with topological signatures. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 1633–1643, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[369] Akihiko Hirata, Tomohide Wada, Ippei Obayashi, and Yasuaki Hiraoka. Structural changes during glass formation extracted by computational homology with machine learning. *Communications Materials*, 1(1):98, 2020.

[370] Paula Carracedo-Reboredo, Jose Liñares-Blanco, Nereida Rodríguez-Fernández, Francisco Cedrón, Francisco J Novoa, Adrian Carballal, Victor Maojo, Alejandro Pazos, and Carlos Fernandez-Lozano. A review on machine learning approaches and trends in drug discovery. *Computational and Structural Biotechnology Journal*, 19:4538–4558, 2021.

[371] Karl G Hill, Danielle Woodward, Tiffany Woelfel, J David Hawkins, and Sara Green. Planning for long-term follow-up: Strategies learned from longitudinal studies. *Prevention Science*, 17:806–818, 2016.

[372] Paweł Dłotko and Hubert Wagner. Simplification of complexes for persistent homology computations. *Homology, Homotopy and Applications*, 16(1):49–63, 2014.

[373] Robert M Samstein, Chirag Krishna, Xiaoxiao Ma, Xin Pei, Ken-Wing Lee, Vladimir Makarov, Fengshen Kuo, Jonathan Chung, Raghvendra M Srivastava, and Tanaya A Purohit. Mutations in brca1. *Nature Cancer*, 1(12):1188–1203, 2020.

[374] Aatish Thennavan, Francisco Beca, Youli Xia, Susana Garcia-Recio, Kimberly Allison, Laura C Collins, M Tse Gary, Yunn-Yi Chen, Stuart J Schnitt, and Katherine A Hoadley. Molecular analysis of tcga breast cancer histologic types. *Cell Genomics*, 1(3):100067, 2021.

[375] Frederick M Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I Olopade, and Jakob N Kather. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12(1):4423, 2021.

[376] Lianli Yin and Yinghua Tang. Predicting prognosis and clinical features of the tumor microenvironment based on ferroptosis score in patients with breast cancer. *Scientific Reports*, 12(1):10611, 2022.

[377] David JH Shih, Naema Nayyar, Ivanna Bihun, Ibiayi Dagogo-Jack, Corey M Gill, Elisa Aquilanti, Mia Bertalan, Alexander Kaplan, Megan R D'Andrea, and Ugonma Chukwueke. Genomic characterization of human brain metastases identifies drivers of metastatic lung adenocarcinoma. *Nature Genetics*, 52(4): 371–377, 2020.

[378] Brian Lane, Mairah T Khan, Ananya Choudhury, Ahmed Salem, and Catharine ML West. Development and validation of a hypoxia-associated signature for lung adenocarcinoma. *Scientific Reports*, 12(1):1290, 2022.

[379] Weimei Huang, Anqi Lin, Peng Luo, Yuchen Liu, Wentao Xu, Weiliang Zhu, Ting Wei, Qingwen Lyu, Linlang Guo, and Jian Zhang. Epha5 mutation predicts the durable clinical benefit of immune checkpoint inhibitors in patients with lung adenocarcinoma. *Cancer Gene Therapy*, 28(7-8):864–874, 2021.

[380] Javad Noorbakhsh, Saman Farahmand, Ali Foroughi Pour, Sandeep Namburi, Dennis Caruana, David Rimm, Mohammad Soltanieh-Ha, Kourosh Zarringhalam, and Jeffrey H Chuang. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nature Communications*, 11 (1):6367, 2020.

[381] Wensheng Zhang and Kun Zhang. A transcriptomic signature for prostate cancer relapse prediction identified from the differentially expressed genes between tp53 mutant and wild-type tumors. *Scientific Reports*, 12(1):10561, 2022.

[382] Philippe Weitz, Yinxi Wang, Kimmo Kartasalo, Lars Egevad, Johan Lindberg, Henrik Grönberg, Martin Eklund, and Mattias Rantalainen. Transcriptome-wide prediction of prostate cancer gene expression from histopathology images using co-expression-based convolutional neural networks. *Bioinformatics*, 38(13): 3462–3469, 2022.

[383] Ashley K Wiseman, Rochelle L Tiedemann, Huihui Fan, Hui Shen, Zachary Madaj, Michael T McCabe, Melissa B Pappalardi, and Peter A Jones. Chromosome-specific retention of cancer-associated dna hypermethylation following pharmacological inhibition of dnmt1. *Communications Biology*, 5(1):528, 2022.

[384] Xingyu Li, Jitendra Jonnagaddala, Min Cen, Hong Zhang, and Steven Xu. Colorectal cancer survival prediction using deep distribution based multiple-instance learning. *Entropy*, 24(11):1669, 2022.

[385] Hyun-Jong Jang, In Hye Song, and Sung Hak Lee. Generalizability of deep learning system for the pathologic diagnosis of various cancers. *Applied Sciences*, 11(2):808, 2021.

[386] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

Appendix A

Additional Embodiment's

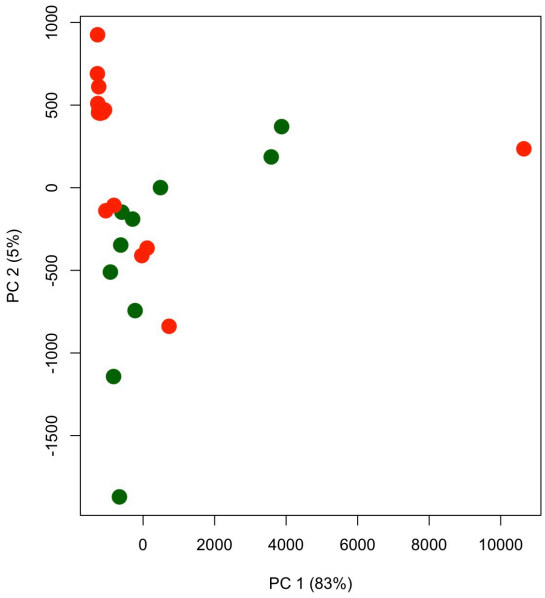
This section presents the results of gene expression and WSI data preprocessing. Additionally, datasets sourced from TCGA for BRCA, LUAD, PRAD, and COAD/READ were analysed and incorporated. The correlation between the signed-TOM and distance correlation measures, along with the subsequently generated topological summaries, was compared across these datasets. Using the same framework with the MLP model for the prediction task of FOLFOX response in CRC patients, the model was applied to classify healthy individuals and cancer patients within the TCGA datasets. The validity of *GeTopology* was also assessed using these additional datasets.

Data Processing

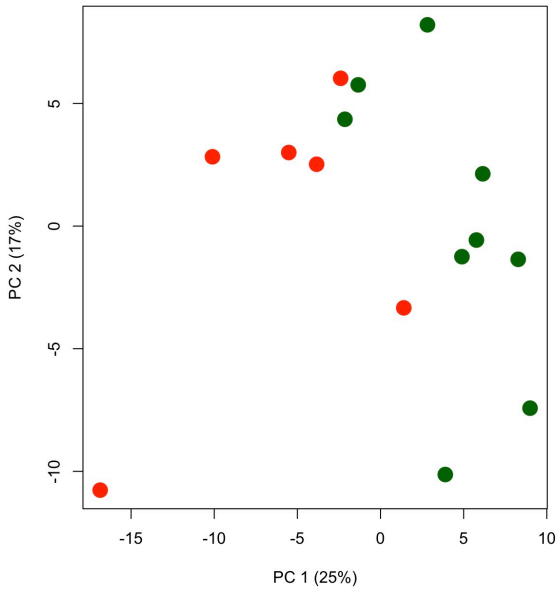
Quality Control

To ensure the robustness of our gene expression data, we implemented rigorous QC procedures for both RNA-Seq and microarray technologies. This involved a thorough examination of sequencing metrics, including read quality and distribution, as well as array-specific QC metrics. Through these procedures, we ensured that the data from both platforms adhered to strict quality standards. The comprehensive QC process significantly improves the reliability of our analyses, confirming that the observed patterns reflect genuine biological signals rather than technical artefacts. To assess sample relationships and the distribution of gene expression quantifications, we employed various methods, including principal component analysis (PCA) (Figure A.1), box-and-whisker plots (box-plots) (Figure A.2), and density plots (Figure A.3). These approaches provided valuable information on the overall quality and integrity of our

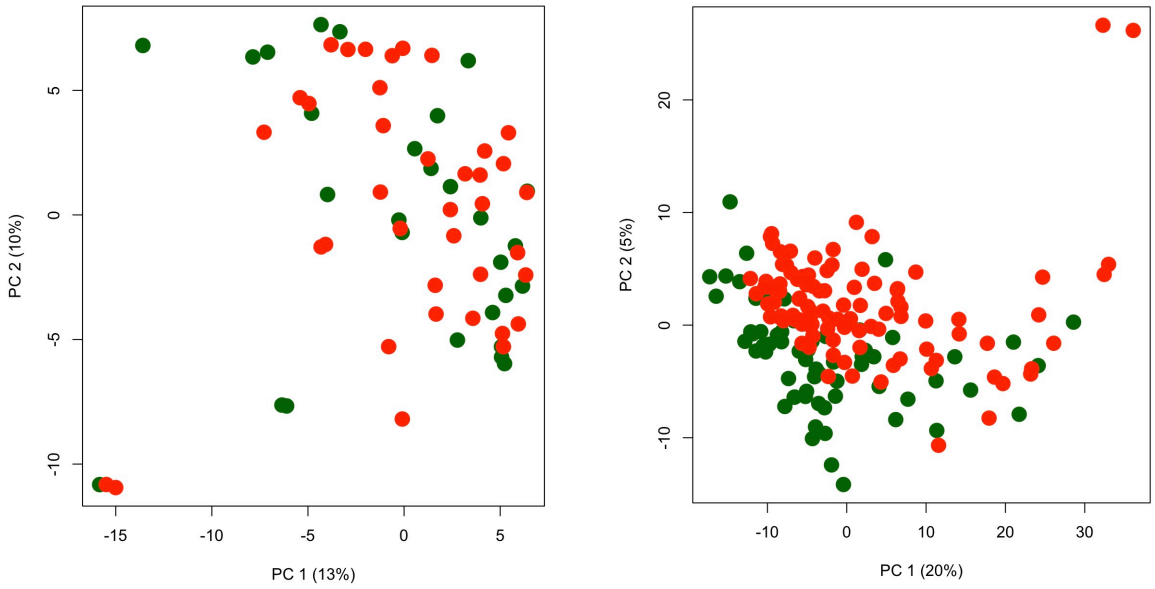
gene expression data. The integration of robust QC measures and diverse assessment methods not only contributes to the credibility of our study but also supports the meaningful interpretation of observed patterns in gene expression.



RNA-Seq

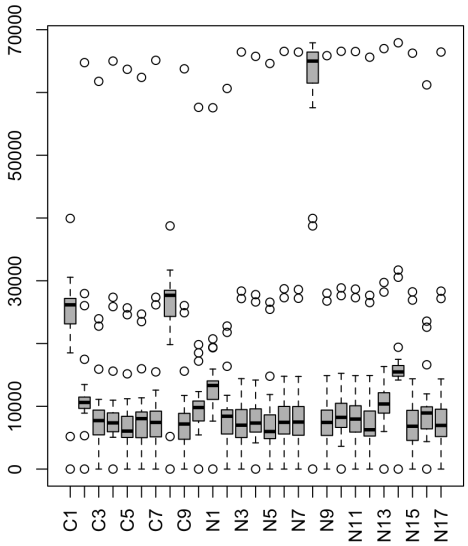


GSE69657

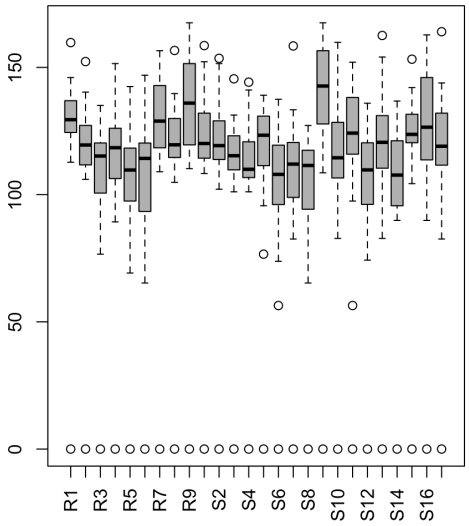


GSE72968 GSE81653

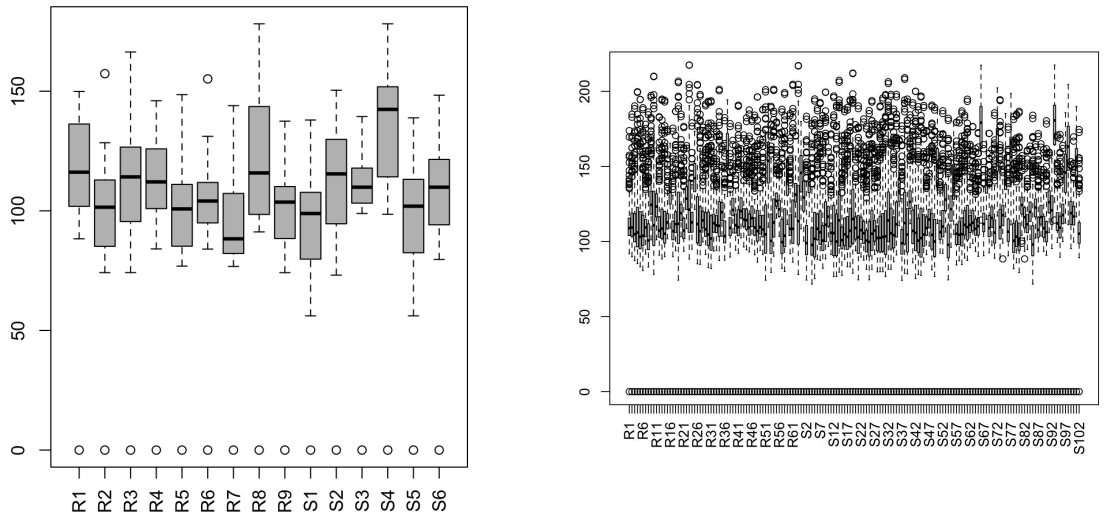
Fig. A.1 The PCA plot illustrates the principal components (PC1 and PC2) for CRC patients. Each dot on the plot represents an individual patient, with blue dots representing patients with CRC resistant to FOLFOX treatment and red dots indicating CRC patients sensitive to FOLFOX. The percentage of variance captured by each axis is depicted, providing insight into the contribution of each principal component to the overall variability in the data. Different colour coding allows for a clear visual differentiation between patients with different responses to FOLFOX treatment, helping to identify potential patterns or clusters in the data. Specifically, PC1 shows a substantial variance of 83% indicative of an outlier for the RNA-Seq dataset



RNA-Seq

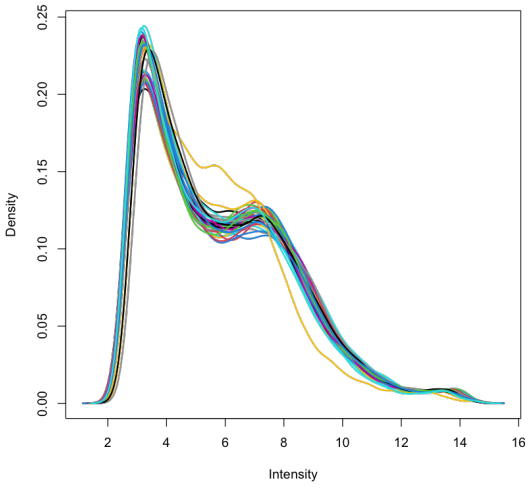


GSE69657

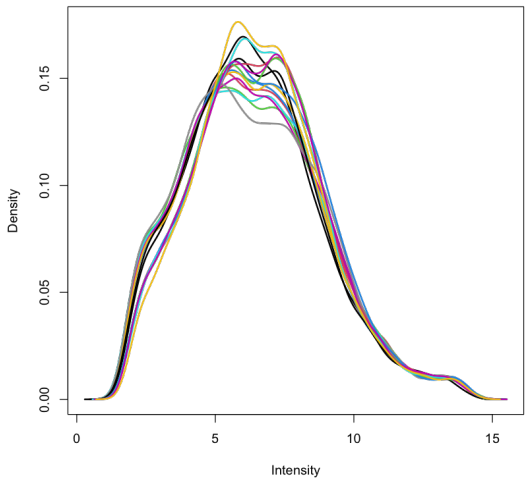


GSE72968 GSE81653

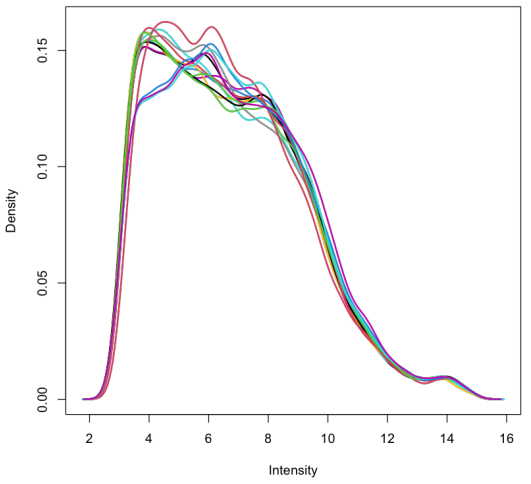
Fig. A.2 Box plots are employed to assess the distribution of data for each sample. On the x-axis, the FOLFOX resistance (R) and sensitivity (S) samples are delineated, while the gene expression measures are depicted on the y-axis. Box plots provide information on the minimum, maximum, and mean values (represented by a horizontal line inside the box). Furthermore, the circles above and below each graph signify values that fall outside the normal range. Further exploration of the outlier in the RNA-Seq dataset is warranted.



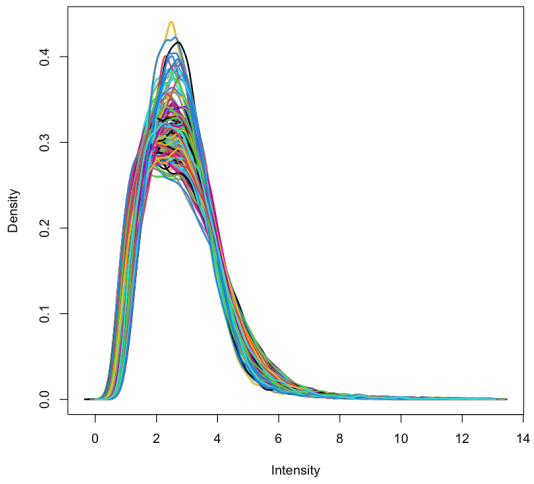
RNA-Seq



GSE69657



GSE72968

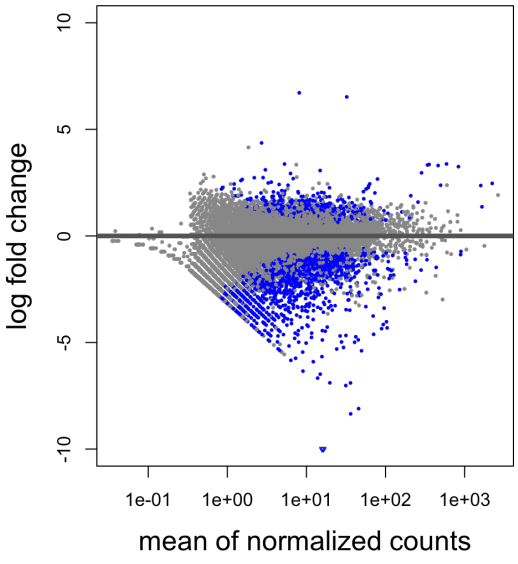


GSE81653

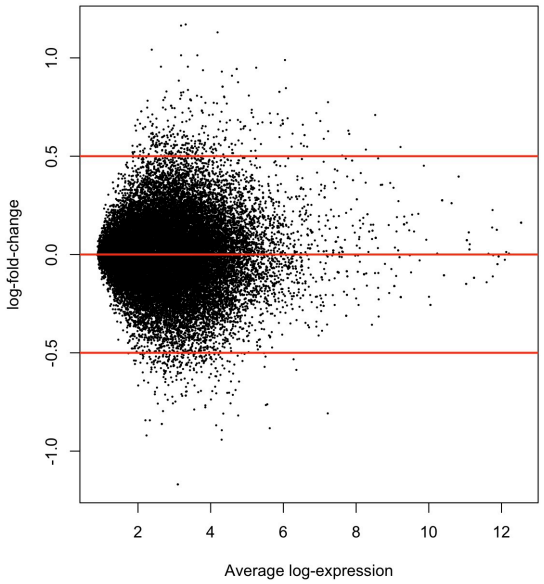
Fig. A.3 The x-axis of the graph represents gene expression values, while the y-axis indicates the density of the concentration of these values. Each sample is distinguished by a unique colour, and the distribution of the data is visually assessed through this representation. Notably, for the RNA-Seq dataset, the presence of an outlying sample represented by the distinctive yellow line is evident. To ensure the robustness and accuracy of downstream bioinformatics and data science tasks, warranted the removal of this sample (healthy individual, N9 - from the Figure A.2).

Differential Gene Expression Analysis

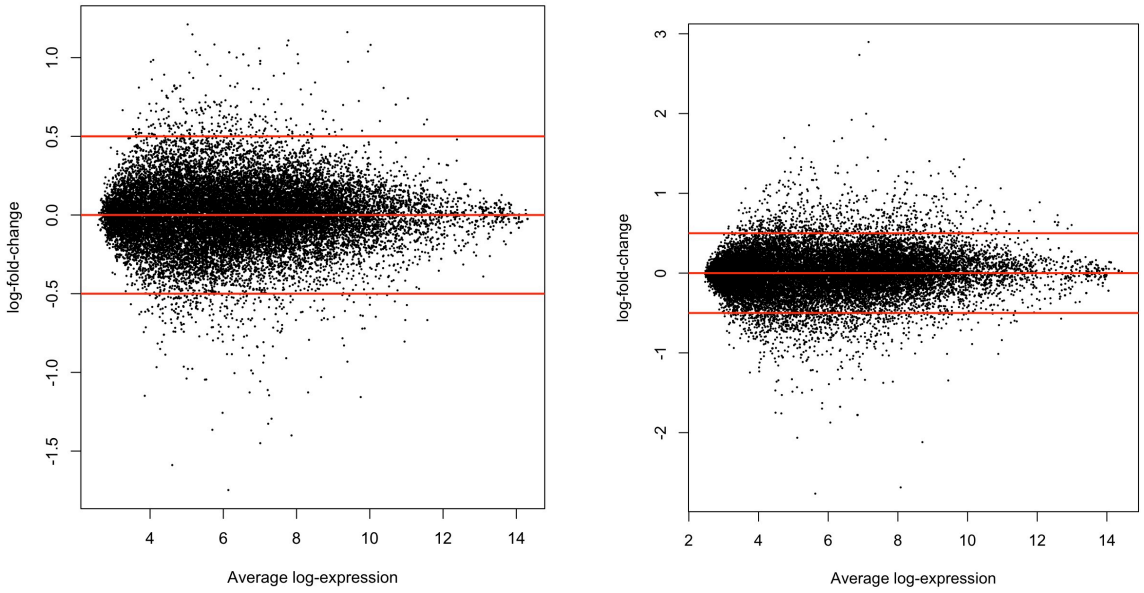
Details of DEGs (Table 3.1) for CRC patients who are resistant or sensitive to FOLFOX treatment can be found in the following GitHub repository link, https://github.com/lebomashatola/GeTopology/tree/main/Differential%20Gene%20Expression. This information was obtained from microarray datasets (GSE81653, GSE69657, GSE72968) and an RNA-Seq dataset. Furthermore, to examine the distribution of genes significantly up- and down-regulated in each gene expression dataset, the MA plot is used. The plot is generated by plotting the average change of log intensities (A) against the differences in the average intensities (M) (Figure A.4). In the legend, A represents the average change in log intensities, M signifies the differences in mean intensities, and the MA plot visually illustrates the distribution of gene expression changes across the dataset.



RNA-Seq



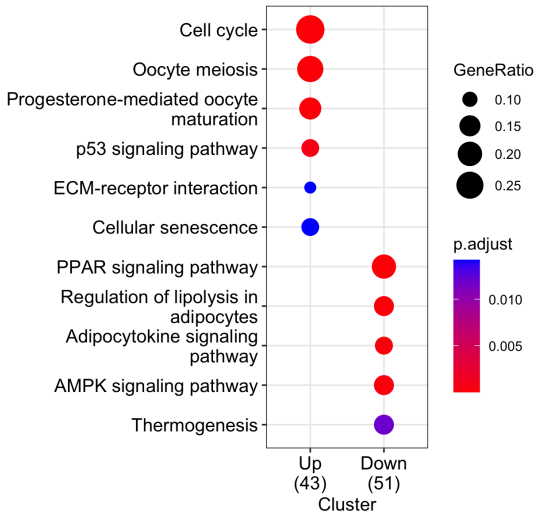
GSE69657



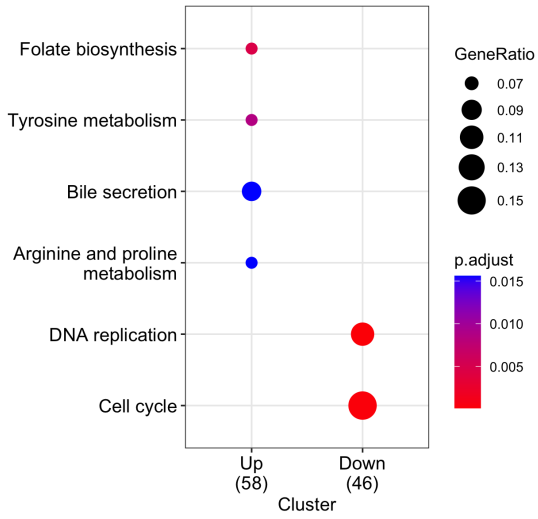
GSE72968 GSE81653

Fig. A.4 MA plots to visually depict the log ratio between gene expression levels in CRC patients classified as FOLFOX-resistant or FOLFOX-sensitive. The x-axis represents the mean expression level, with thresholds set at 0.5 and -0.5 to enhance the identification of DEGs. The x-axis is tailored for each dataset, displaying the average log expression for microarray datasets and mean-normalised counts for the RNA-Seq dataset. The y-axis consistently represents log fold change, ensuring a standardised and comparable analysis of differential expression. For RNA-Seq dataset, the grey points represent non-significant DEGs, whilst blue points are significant DEGs.

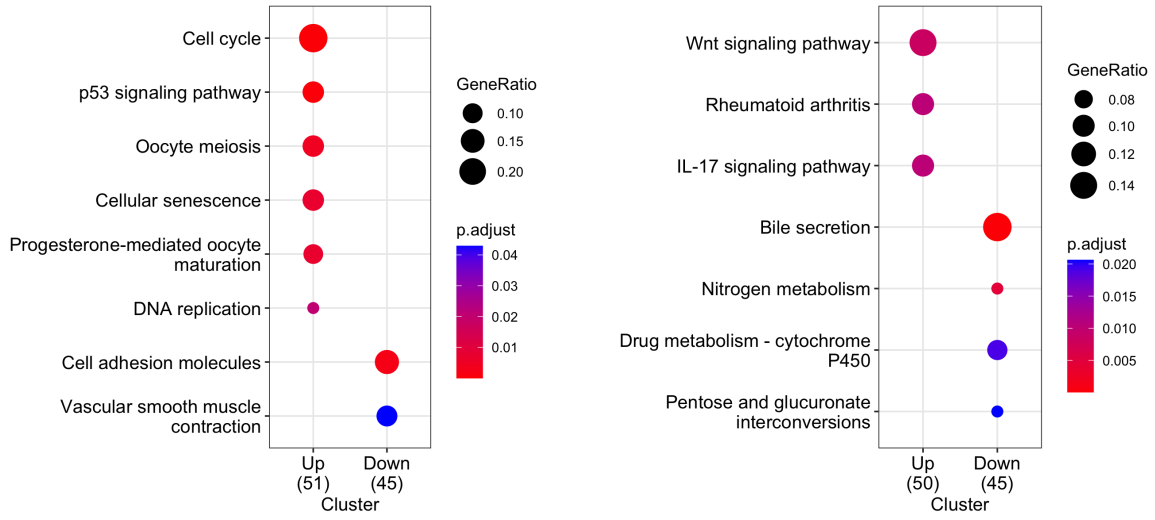
Table 3.1 summarises DEGs in various datasets, including GSE69657, GSE81653, GSE72968, and RNA-Seq. It outlines the counts of both up-regulated and down-regulated genes in each dataset. For example, GSE69657 has 123 up-regulated and 115 down-regulated genes, GSE81653 exhibits 230 up-regulated and 210 down-regulated genes, GSE72968 shows 119 up-regulated and 113 down-regulated genes, and the RNA-Seq dataset displays 150 up-regulated and 210 down-regulated genes. This concise presentation provides an overview of the unique patterns of differential gene expression in each dataset, providing insight into the regulatory dynamics of genes within their specific contexts. The exploration of resistance to FOLFOX involved a crucial step of functional enrichment analysis to reveal over-represented biological annotations within the DEGs. Initially identified by comparing gene expression profiles between FOLFOX resistant and FOLFOX sensitive conditions, DEGs were functionally enriched using databases including the KEGG pathway databases. The results, which illustrate the enriched biological annotations, played a key role in interpreting the molecular foundations of resistance to FOLFOX (Figure A.5).



GSE81653



GSE69657

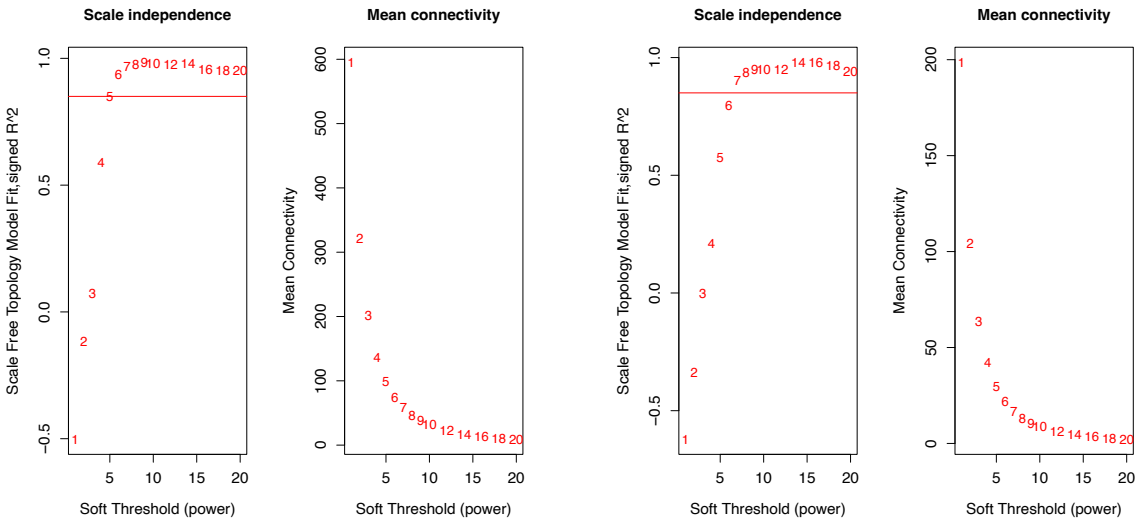


GSE72968 RNA-Seq

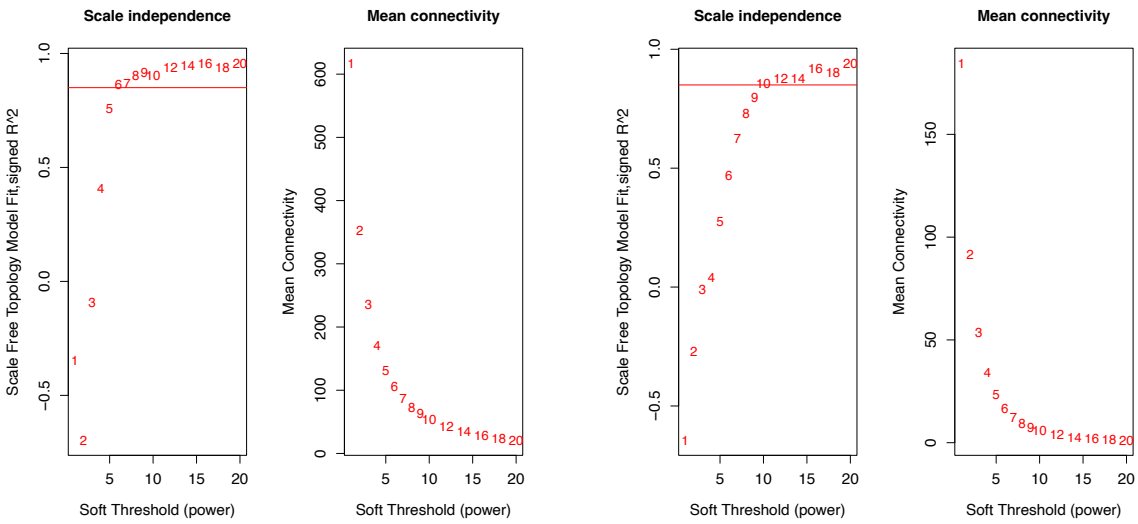
Fig. A.5 The figure represents the results of the analysis of gene module enrichment in DEGs. The x-axis represents the number of genes over-represented in KEGG for each gene module, while the y-axis illustrates the enrichment of KEGG terms. Each point in the graph corresponds to a specific KEGG term associated with a gene module. The colour of the points indicates the BH-adjusted p-value, with lower values indicating higher statistical significance. Additionally, the size of the points reflects the gene ratio, representing the percentage of total co-expressed genes within the given KEGG term. Larger points indicate a higher gene ratio, highlighting the proportion of genes from the input data set that correspond to the enriched KEGG pathway. This comprehensive visualisation provides insights into both the statistical significance and biological relevance of gene module enrichment in the analysed dataset. Notable findings include the upregulation of the p53 signalling pathway and cell cycle in GSE81653 and GSE72968 datasets, the upregulation of the Wnt signalling pathway in the RNA-Seq dataset, and the downregulation of DNA replication in GSE72968 and GSE69657 datasets. These specific enrichments underscore the biological relevance of the gene modules and provide insights into the potential dysregulation of key pathways in CRC. The comprehensive visualisation the understanding of both the statistical significance and biological implications of gene module enrichment in the analysed dataset.

Weighted Adjacency Matrix

The determination of the soft threshold parameter, denoted as β, is a crucial step in the transformation of the weighted adjacency matrix in network analysis, particularly in the context of co-expression analysis. This parameter significantly influences the construction of the adjacency matrix, representing the strength of connections between nodes in a network. In this analysis, the selection of the lowest β value for each dataset was pivotal for transforming the pre-selected gene expression matrix, using the identified DEGs, into a meaningful weighted adjacency matrix. This careful choice of an appropriate β is essential to capture biologically relevant relationships among genes and optimise the construction of the weighted adjacency matrix. Importantly, the gene expression data pre-selected using DEGs served as the input for this analysis. This optimisation ultimately enhances the accuracy of subsequent network-based analyses, providing a solid foundation for exploring co-expression patterns and identifying key gene modules associated with resistance to FOLFOX treatment.



RNA-Seq GSE69657



GSE72968 GSE81653

Fig. A.6 The soft-thresholding values were evaluated to identify the optimal R^2 and mean connectivity. Plots showing scale independence and mean connectivity across soft-thresholding values ranging from one to twenty were generated. The optimal R^2 and its corresponding β value were determined based on the scale-independent plot, while the optimal mean connectivity and its associated β value were identified from the mean connectivity plot. Notably, the dataset GSE81653, exhibited the least association with a scale-free topology. Larger soft-thresholding values, reaching an R^2 of 0.85, were observed for this dataset. This unique characteristic of GSE81653 highlights its distinct network structure compared to other datasets.

Hub Gene Selection

In the process of identifying the hub genes, we calculated intramodular connectivity, which assesses the strength of gene-gene interactions within a specific gene module. Hub genes were defined as the top 15% of highly connected genes within each gene module. To establish a connection strength threshold for selecting hub genes within each module, we evaluated intramodular connectivity and module membership relationships. This approach accounted for the variability observed between gene modules, as depicted in Figure A.7. This step was crucial in filtering and pinpointing hub genes based on their connectivity patterns within distinct gene modules.

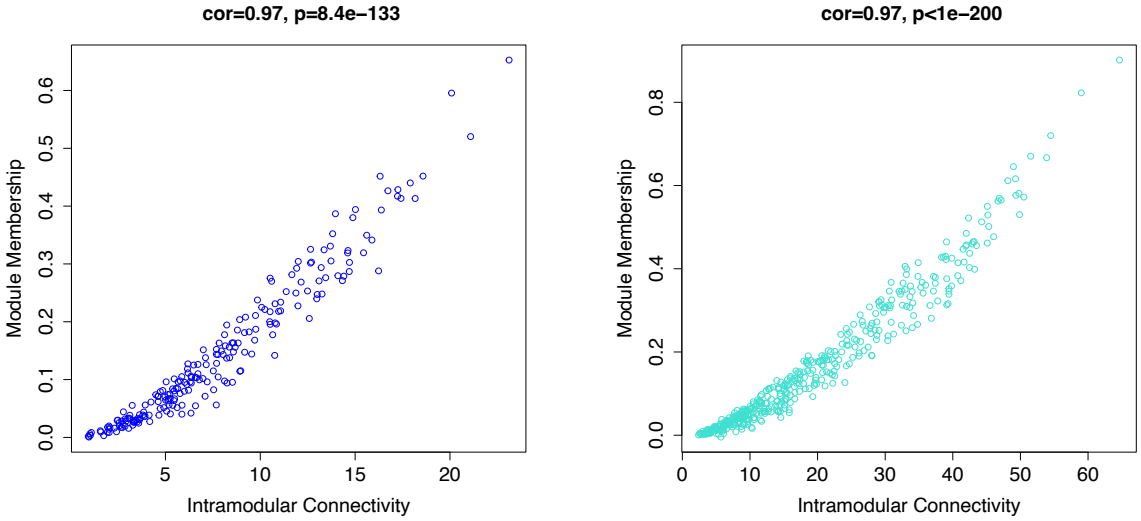
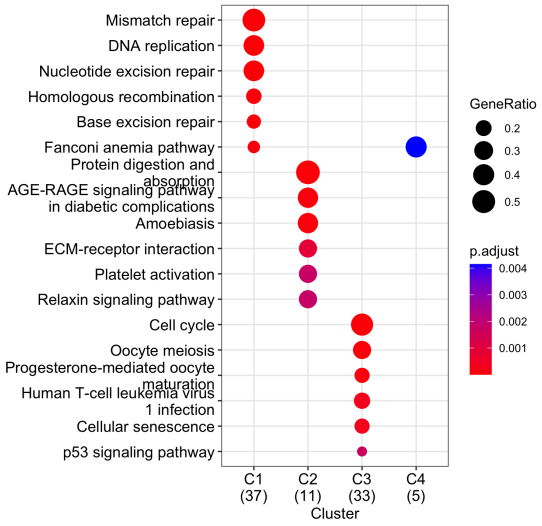


Fig. A.7 A graph is presented which plots the intramodular connectivity of genes in a gene module against their module membership (k_{ME}). The data points demonstrate that there is a range of maximal intramodular connectivity and module membership, suggesting that a constant k_{ME} threshold should not be used to select hub genes for each gene module. The x-axis of the graph displays the intramodular connectivity and the y-axis shows the module membership (k_{ME}). Furthmore, the Pearson correlation coefficient is shown by "corr" and "p" indicates the p-value to assess the strength and significance of the relationship between a gene and its assigned gene module.

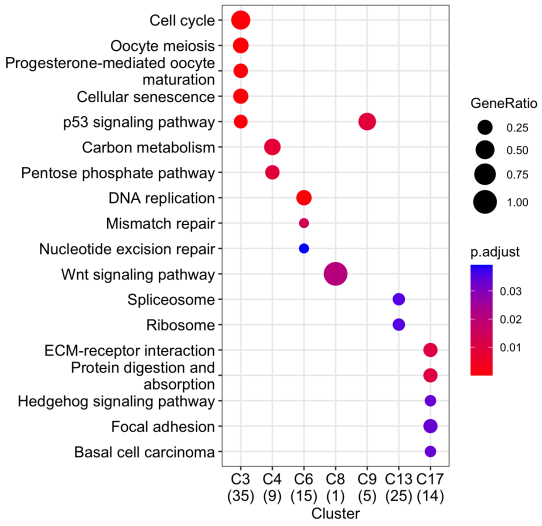
Biological Annotation of Hub Genes

In the results section concerning functional enrichment analysis, the implementation to identify over-represented KEGG pathways, aiming to enrich our understanding of

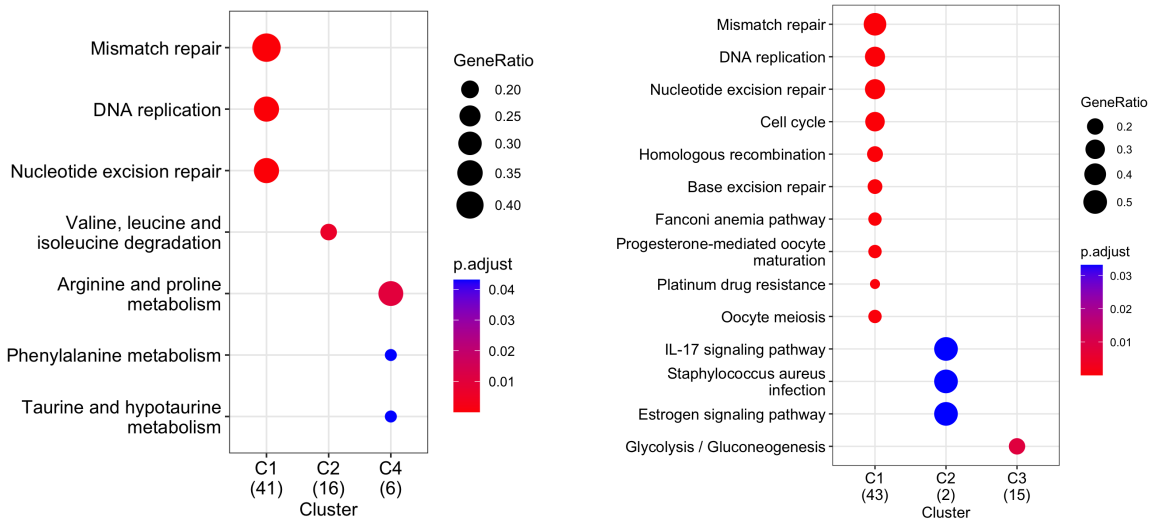
the biological pathways associated with hub genes in patients with CRC resistant to FOLFOX treatment. The hub genes were identified from the gene expression matrix pre-selected using DEGs. Subsequently, k_{ME} values were determined, and appropriate thresholds were set to select the top 15% most connected genes in each gene module. Finally, these hub genes were annotated using the KEGG pathway database for over-representation analysis. Figure A.8 illustrates the results of this comprehensive analysis.



GSE81653



GSE69657



GSE72968 RNA-Seq

Fig. A.8 The *x-axis* represents the quantity of hub genes within each module (labelled "C" and the corresponding gene module number), and the *y-axis* illustrate the enriched KEGG pathway. Point colour indicates the BH-adjusted p-value, with a colour gradient representing the significance level. Point size corresponds to the gene ratio, which is the proportion of total hub genes associated with the respective KEGG term. Gene ratio is defined as the number of mapped genes to the specific KEGG pathway from the input dataset. Larger points indicate a higher gene ratio. The selection of significantly enriched KEGG pathways is based on a stringent criterion of BH-adjusted p-value less than 5%, ensuring robust and statistically significant functional annotations associated with the hub genes.

Topological Summaries

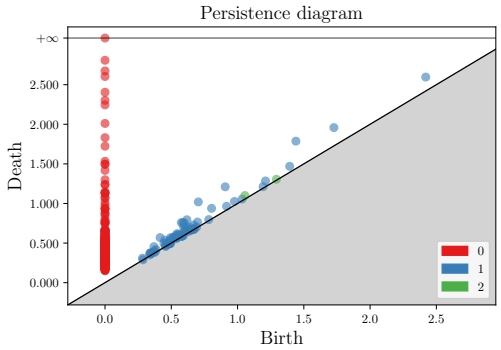
The results compare gene expression matrices transformed using signed-TOM to distance correlation. To validate new discoveries, the signed-TOM transformations of gene expression data were utilised to improve the identification of topological features. The precision of the prediction model using topological summaries in the form of persistence images was evaluated on TCGA-BRCA, TCGA-COAD/READ, TCGA-LUAD, and TCGA-PRAD datasets. Table A.1 summarises the number of cancer samples for each phenotype. Furthermore, to further assess the performance and generalisability of the model, independent datasets were included in this investigation due to the limited number of patients with CRC resistant/sensitive to FOLFOX.

Table A.1 Summarising information on cancer datasets obtained from The Cancer Genome Atlas (TCGA). The table abbreviates each dataset, specifying breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), prostate adenocarcinoma (PRAD), and colorectal adenocarcinoma/read (COAD/READ). For each dataset, the table indicates the number of samples classified as cancer-afflicted and healthy, giving insight into the composition of the dataset. The reference column cites studies associated with each dataset, providing a comprehensive understanding of the dataset sources and research context. Notable references for previous work performed using the cohorts include: BRCA [373–376], LUAD [201, 377–380], PRAD [381, 382], and COAD/READ [383–385].

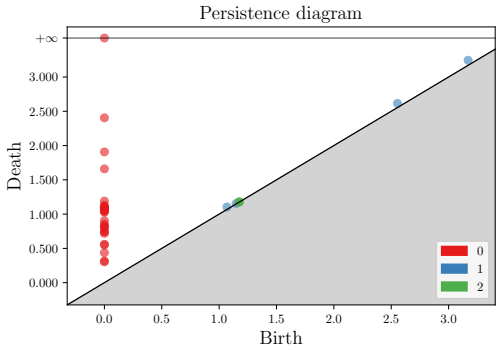
Dataset	**Cancer-afflicted**	**Healthy**	**References**
BRCA	193	132	[373–376]
LUAD	132	107	[201, 377–380]
PRAD	115	51	[381, 382]
COAD/READ	123	50	[383–385]

In the results section, PDs were used to visually represent the topological features resulting from both signed-TOM and distance correlation transformations. The use of PDs allows for a comprehensive visualisation of persistent homological features in the data, providing insight into the persistence of topological structures. Figure A.9 serves as a representative example, illustrating a persistent diagram generated for a single patient. Each point in the diagram corresponds to a topological feature, and the position of the points reflects the birth and death of these features throughout the persistence-based filtration process. Analysing PDs aids in discerning unique topological patterns and understanding the persistence of these patterns across different transformations, contributing to a more nuanced interpretation of the data's structural

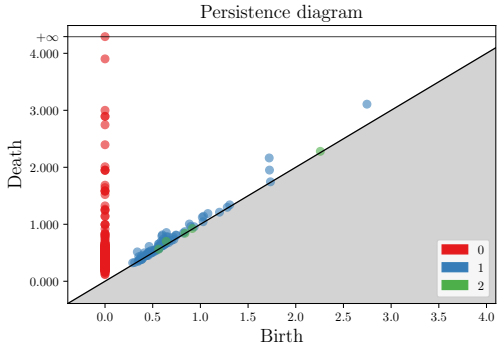
characteristics.



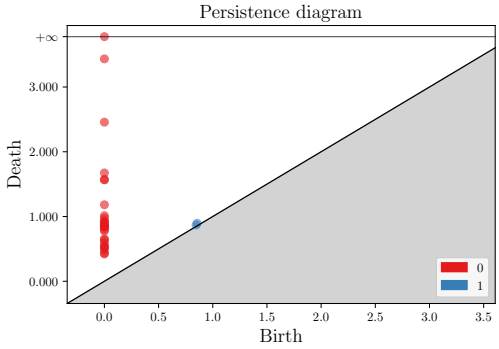
BRCA - Signed-TOM



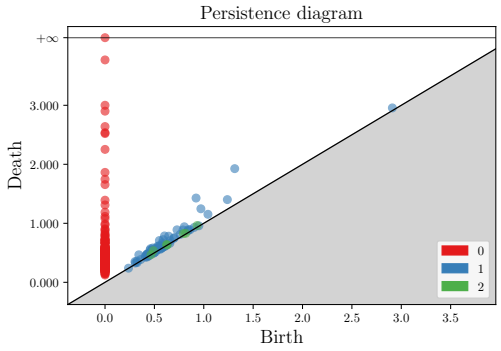
BRCA - Distance Correlation



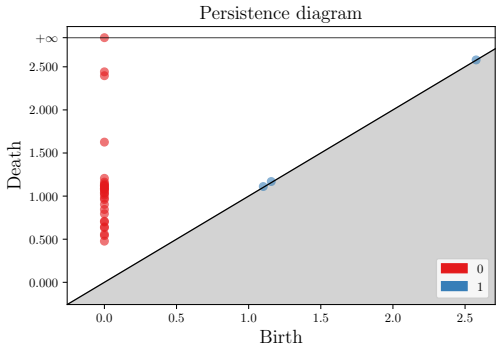
LUAD - Signed-TOM



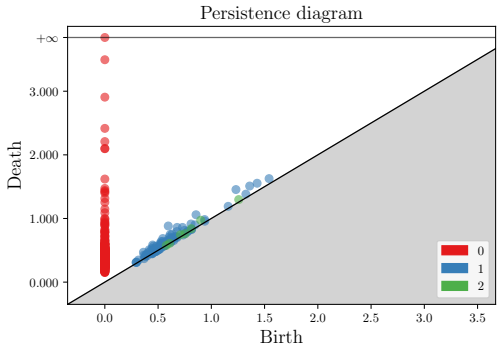
LUAD - Distance Correlation



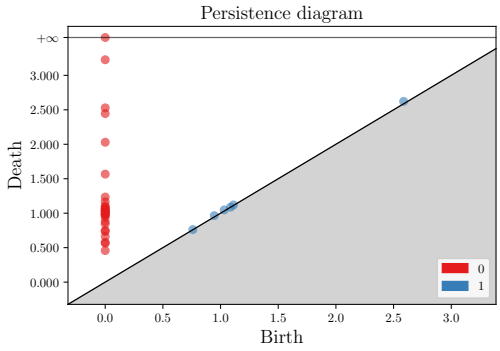
PRAD - Signed-TOM



PRAD - Distance Correlation



COAD-READ - Signed-TOM



COAD-READ - Distance Correlation

Fig. A.9 Randomly selected patient's PD for independent Datasets for a randomly selected cancer patient's created using signed-TOM measurements and distance correlation measures. The x-axis represents the birth parameter (b), and the y-axis represents the death parameter (d). Topological features, classified into homology groups, are depicted as points: *Betti*-0 (in red), *Betti*-1 (in blue), and *Betti*-2 (in green). The grey area denotes regions where no topological features were identified (i.e., where b is less than d). Notably, the signed-TOM (left) illustrates more topological features compared to the distance correlation (right). This observation suggests that the signed-TOM method captures a more intricate and detailed representation of the patient's persistence in the data, providing richer insights into the topological structure associated with cancer patient characteristics.

Prediction Learning

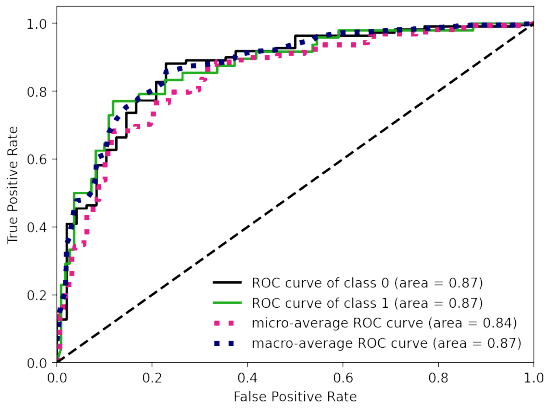
Table A.2 summarises the performance metrics of the model obtained from the classification of patients as cancerous or healthy according to their topological signatures. Signatures were generated for each patient using an MLP model. The classification was performed using two different approaches: distance correlation and signed-TOM. The table includes key performance metrics that provide information on the accuracy and effectiveness of the models in distinguishing between cancer patients and healthy individuals.

Table A.2 An evaluation of the precision of the prediction of phenotypes in validation data using multi-layer perceptron (MLP) models. The accuracy values are presented as a mean with a standard deviation (±). The values in parentheses represent the corresponding sensitivity (true positive rate) for each dataset. The analysis involves two transformation methods: Distance correlation and signed-topological overlap matrix (Signed-TOM). The datasets include Breast Invasive Carcinoma (BRCA), Lung Adenocarcinoma (LUAD), Colorectal Adenocarcinoma/Read (COAD/READ), and Prostate Adenocarcinoma (PRAD). Results highlight varying accuracies for different datasets and transformation methods (signed-TOM and distance correlation measures).

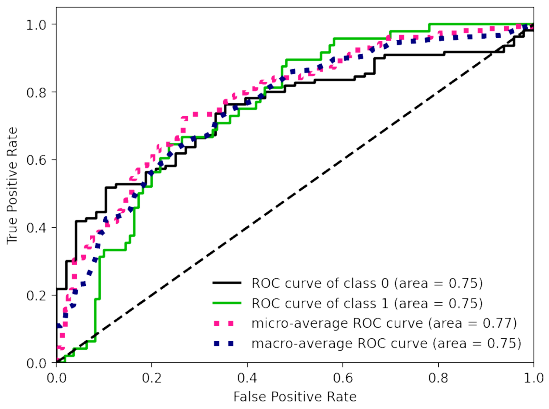
Dataset	**MLP Accuracy**	
	Distance Correlation	**Signed-TOM**
BRCA	76.51 ± 1.78 (78.54 ± 1.45)	87.71 ± 1.48 (91.36 ± 1.46)
LUAD	70.83 ± 1.76 (73.56 ± 1.21)	91.36 ± 0.98 (93.25 ± 1.45)
COAD/READ	72.81 ± 2.90 (74.35 ± 1.99)	96.11 ± 0.06 (96.47 ± 0.81)
PRAD	68.34 ± 0.98 (73.47 ± 1.11)	87.34 ± 0.11 (91.00 ± 0.71)

Note: MLP training accuracy performance is shown in parentheses.

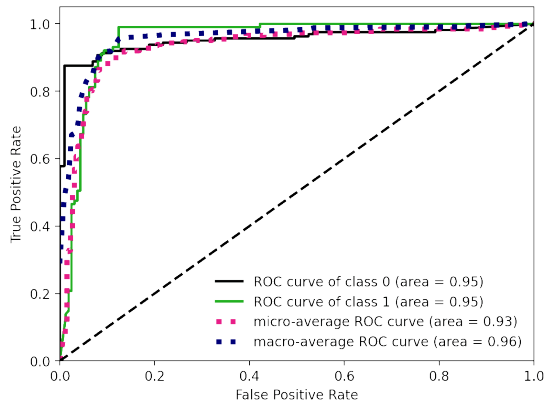
The ROC curves, depicted in Figure A.10, illustrate the performance of the MLP model in classifying patients as having cancer or being healthy. The topological summaries used as input were constructed using either distance correlation measures or signed-TOM. These curves provide a visual representation, providing insights into the model's ability to differentiate between patients with varied responses to FOLFOX treatment based on the topological features derived from the input data.



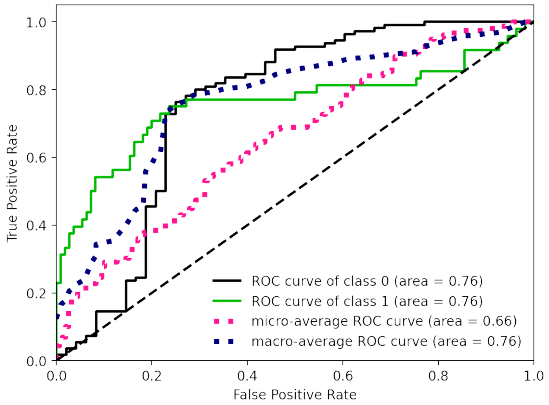
Signed-TOM$_{\text{Test}}$ BRCA



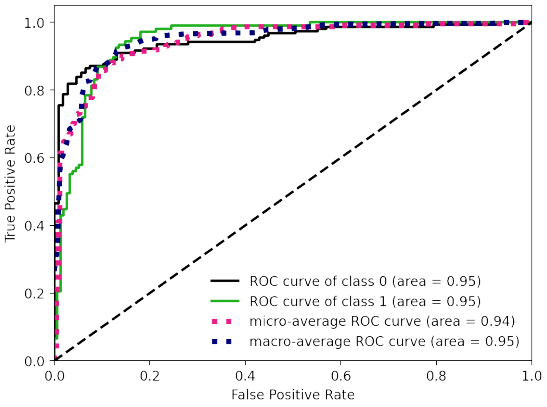
Distance Correlation$_{\text{Test}}$ BRCA



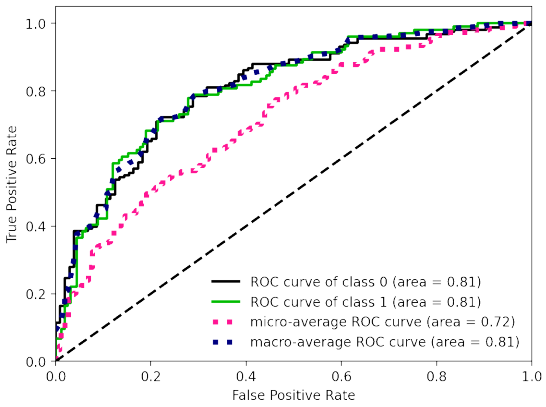
Signed-TOM$_{\text{Test}}$ LUAD



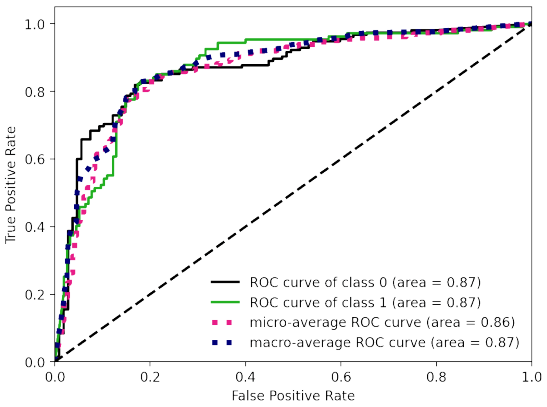
Distance Correlation$_{\text{Test}}$ LUAD



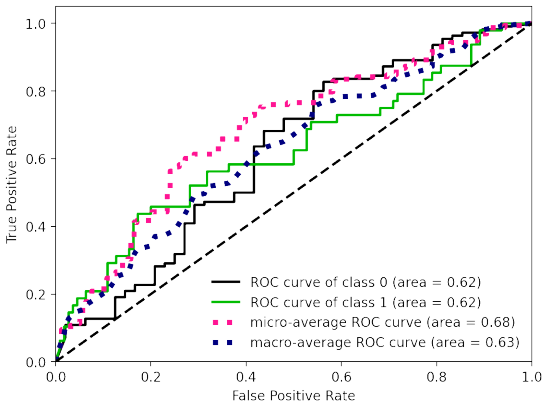
Signed-TOM$_{\text{Test}}$ COAD/READ



Distance Correlation$_{\text{Test}}$ COAD/READ



Signed-TOM$_{\text{Test}}$ PRAD



Distance Correlation$_{\text{Test}}$ PRAD

Fig. A.10 The receiver operating characteristic (ROC) curves compare the performance of signed-TOM and distance correlation measures in distinguishing between cancer patients and healthy individuals for independent datasets. The x-axis represents the FPR, while the y-axis represents the TPR. F1 scores, both micro- and macro-, are reported along with the AUC for each predicted class. These metrics provide insight into the precision, recall, and overall discriminative power of the model in identifying cancer patients and healthy individuals. Notably, Signed-TOM consistently surpasses distance correlation, highlighting its superior performance in phenotype prediction tasks.

Deep Learning on *GeTopology*

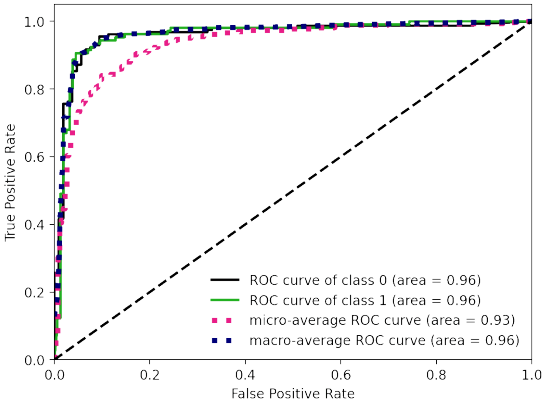
Table A.3 presents the results of a comparative analysis between *GeTopology* and *NSCLC*. Performance metrics were evaluated using independent cancer datasets, providing information on the effectiveness of both models in handling diverse datasets and capturing essential biological information. The key performance indicators of the model are summarised, allowing for a comprehensive comparison between *GeTopology* and *NSCLC* in the context of independent cancer datasets.

Table A.3 Reporting on the comparison of classification performance between *NSCLC* and *GeTopology* models across independent cancer datasets. The table provides accuracy values as a mean and the standard deviation (±) for each model to distinguish cancer phenotypes in breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), colon and rectal adenocarcinoma (COAD/READ), and prostate adenocarcinoma (PRAD). The values in parentheses represent the corresponding model training mean accuracy scores. Notably, *GeTopology* consistently outperforms *NSCLC* in accurately classifying cancer phenotypes across all evaluated datasets, showcasing its superior predictive capabilities.

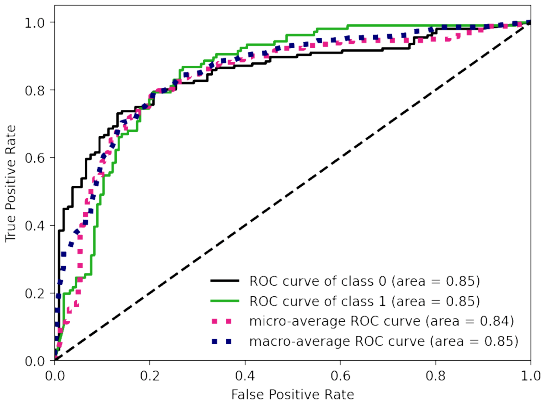
Dataset	*NSCLC*	*GeTopology*
BRCA	84.00 ± 0.61 (87.64 ± 1.35)	95.31 ± 0.25 (97.01 ± 0.36)
LUAD	88.00 ± 0.65 (89.01 ± 0.88)	96.04 ± 0.45 (98.02 ± 0.64)
COAD/READ	87.09 ± 0.76 (89.19 ± 0.91)	96.84 ± 0.90 (97.35 ± 0.95)
PRAD	84.09 ± 0.99 (85.12 ± 1.11)	90.60 ± 1.10 (92.94 ± 1.17)

Note: *GeTopology* and *NSCLC* training accuracy performance is shown in parentheses.

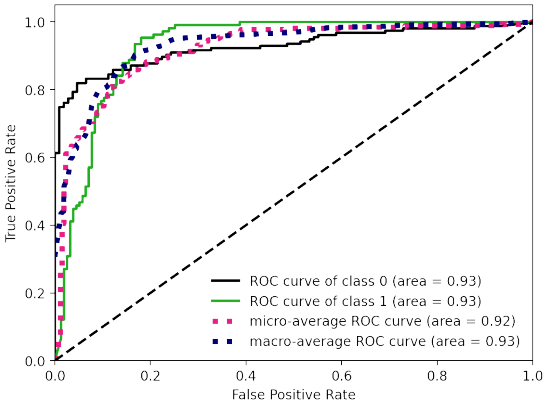
ROC curves, depicted in Figure A.11, provide a visual representation of the classification performance of the MLP model to distinguish between cancer patients and healthy individuals. The topological summaries used as input were generated through distance correlation measures or signed-TOM for independent datasets. These curves offer valuable information on the efficacy of the model in distinguishing between cancer patients and healthy individuals, using the topological features derived from the input data.



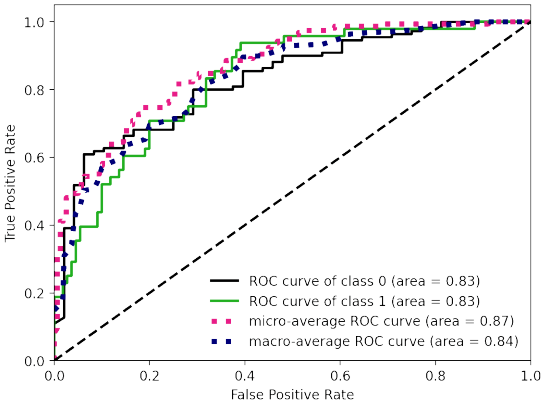
$GeTopology_{\text{Test}}$ BRCA



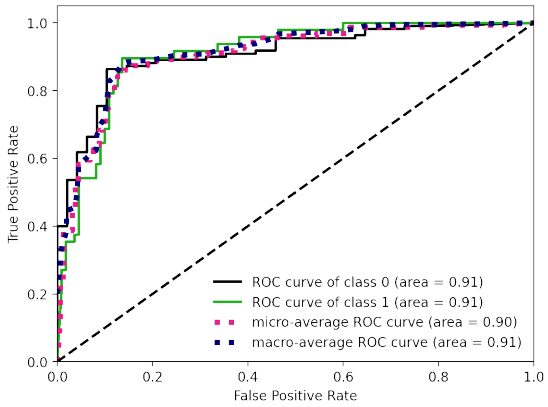
$NCSLC_{\text{Test}}$ BRCA



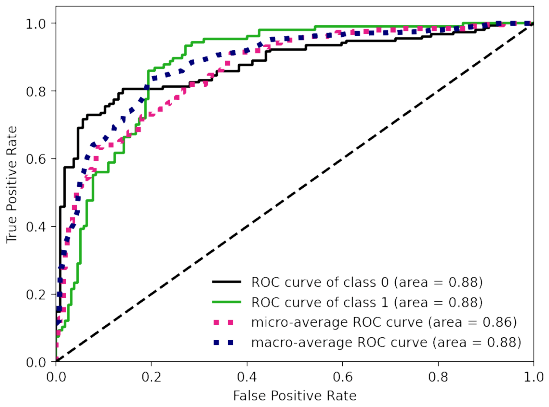
$GeTopology_{\text{Test}}$ LUAD



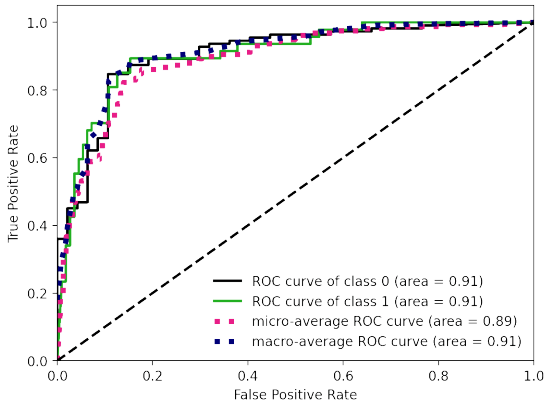
$NCSLC_{\text{Test}}$ LUAD



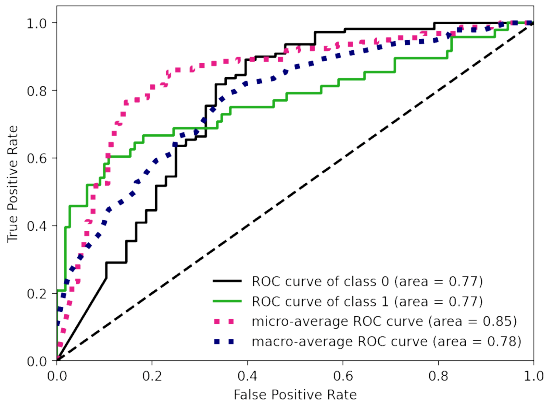
$GeTopology_{\text{Test}}$ COAD/READ



$NCSLC_{\text{Test}}$ COAD/READ



$GeTopology_{\text{Test}}$ PRAD



$NCSLC_{\text{Test}}$ PRAD

Fig. A.11 The receiver operating characteristic (ROC) curves compare the performance of *GeTopology* and *NCSLC* in distinguishing between cancer patients and healthy individuals for independent datasets. The x-axis represents the FPR, while the y-axis represents the TPR. F1 scores, both micro- and macro-, are reported along with the AUC for each predicted class. These metrics provide insight into the precision, recall, and overall discriminative power of the model in identifying cancer patients and healthy individuals. Notably, *GeTopology* consistently outperforms *NSCLC* in accurately classifying cancer phenotypes across all evaluated datasets, showcasing its superior predictive capabilities.

Appendix B

Running the *GeTopology* Framework

To illustrate the functionality of the framework, we provide a demonstration on Unix systems. The framework incorporates advanced features such as the processing of paired-end *fastq* data for RNA-Seq. Key processing steps include quality control, alignment with the reference genome, transcript annotation, and quantification. Table B.1 outlines the computational tools used by *GeTopology*.

Before using the *GeTopology* framework, clone the Github repository on your system (https://github.com/lebomashatola/GeTopology). Ensure that the necessary software is installed on your computer by examining the *Requirements.txt* file at https://github.com/lebomashatola/GeTopology/blob/main/requirements.txt, which contains all the Python packages needed, and by executing the R programming command to install the necessary R programming packages.

```
Rscript requirements.r
```

Downloading Data

Retrieve cancer gene expression data from the https://portal.gdc.cancer.gov website. Organise the main folder with a specific label for the phenotype of interest. Within this main folder, create subfolders, each containing HTSeq count files in tab separated values (TSV) format. Ensure a consistent structure for the WSI data, following the

Table B.1 Reporting on the RNA-Seq processing tools employed within the *GeTopology* framework. The table outlines the specific tasks, the corresponding tools, the associated references and URLs for each tool. Tools include *Trimmomatics*, *FastQC*, *Samtools* for quality control (QC), *HISAT2* for Reference Genome Alignment, and *HTSeq* for Transcript Annotation and Quantification. Hyperlinks have been provided for each URL in the table.

Task	**Tool**	**References**	**URL**
QC	*Trimmomatics*, *FastQC*, *Samtools*	[95, 96, 386]	http://www.usadellab.org/cms/?page=trimmomatic https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ https://www.htslib.org/
Reference Genome Alignment	*HISAT2*	[102]	https://daehwankimlab.github.io/hisat2/
Transcript Annotation and Quantification	*HTSeq*	[159]	https://htseq.readthedocs.io/en/latest/

Note: A hyperlink has been added for each URL in the table.

same folder organisation, with images in *SVS* format. Adherence to this directory structure is crucial for the proper functioning of the framework and accurate data processing (Figure B.1).

Data Processing

To perform an analysis of RNA-Seq data in fastq format, the following tasks will be executed: QC of fastq data, trimming of adapter sequences, genome alignment, and quantification at the transcript level. Execute the code provided:

```
python RNA-Seq_Processing.py
```

For processing of microarray data (in CEL format and on the GPL570 platform), the subsequent code is essential. It performs RMA normalisation and merge the files into a comprehensive gene expression matrix. Executing the following code is crucial for effective data processing:

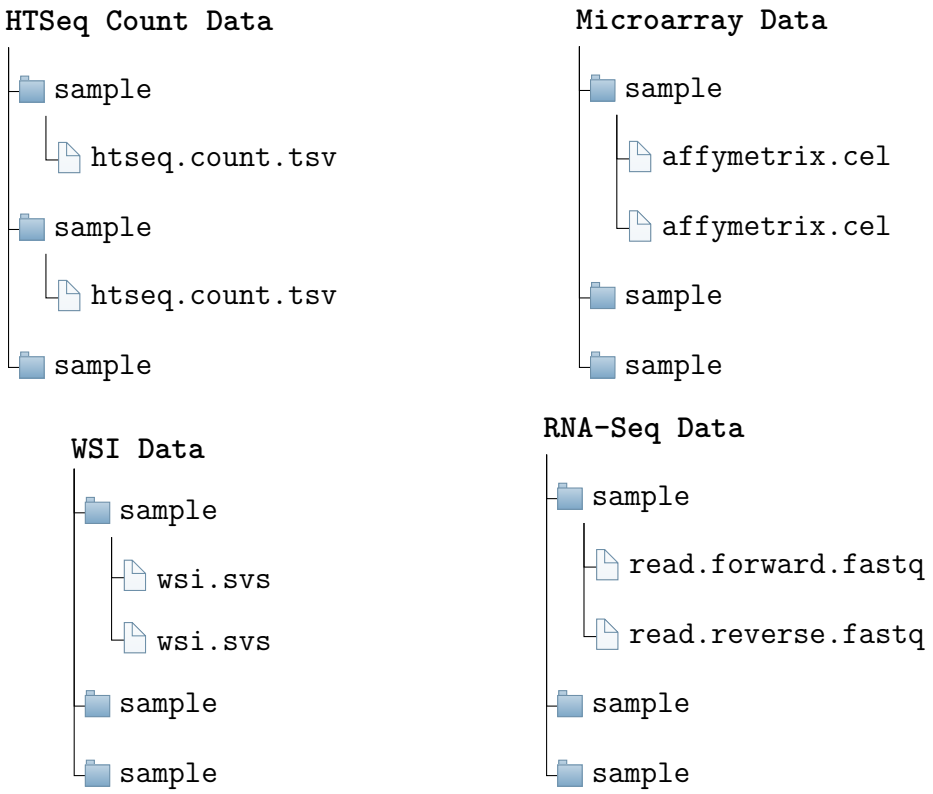


Fig. B.1 The directory tree structure for input data in the *GeTopology* framework is shown. The main folder is labelled with the particular phenotype, and within it, the subfolders contain HTSeq count files (in *TSV* - tab-separated values format) for RNA-Seq data. Alternatively, the subfolders contain microarray data (in CEL - Affymetrix Cell Intensity File format). Additionally, the same structure is applied for WSI data, with images formatted in slide virtual slide (*SVS*) format. This organised arrangement ensures the correct processing of data by the framework, facilitating the integration of RNA-Seq and WSI data for analysis.

```
python run_cels.py
```

To preprocess WSI data, which must be in *SVS* format, the essential steps involve image segmentation and image tiling. The code provided below accomplishes these tasks:

```
python Image_Processing.py
```

In the processing of HTSeq counts obtained from GDC portal, the initial steps involve the selection of protein-coding genes and the amalgamation of individual tab-separated value (*TSV*) files into a comprehensive gene expression matrix. To facilitate this process, execute the provided command:

```
python Expression_Counts.py
```

Prediction Learning of Topological Summaries

The prediction task executed on gene expression data, encompassing both microarray and HTSeq counts, requires a transformation into topological summaries. This process begins with the computation of differential gene expression and the analysis of co-expression to pinpoint hub genes. Following this, the signed-TOM between hub gene pairs within the pre-selected gene expression matrix is calculated. Subsequently, the PH and the vector transformation of the PD into a PI are computed. The PI of each patient is labelled according to the phenotype and classification between two classes is performed using an MLP. The resulting model performance metrics are then outputted, as detailed in Section 2.6.1. To successfully undertake prediction learning using topological summaries, it is imperative to follow these sequential steps:

1. Download the cancer gene expression from https://portal.gdc.cancer.gov.
2. Using both the required R and Python packages are installed.
3. Identify hub genes using *DESeq2* for RNA-Seq or *Limma* for microarray data

```
python run_preselection.py
```

4. Generate PI labelled for each patient

```
python generate_persistent_diagrams_images.py
```

5. Perform classification using the PIs

```
python phenotype_prediction.py
```

Classification of Whole-Slide Images

The initial processing of image data involves image segmentation and image tiling. Subsequently, a MobileNet is used to classify between two distinct phenotypes.

1. Download the WSI for cancer patients from https://portal.gdc.cancer.gov
2. Image processing

```
python Image_Processing.py
```

3. Image classification

```
python MobileNet.py
```

Deep Learning using Bimodalities

To undertake DL using bimodalities, it is essential to ensure that the directory tree aligns correctly with the structure illustrated in Figure B.2.

1. Run the DL model using bimodalities for the binary classification of cancer phenotypes.

```
python integrate.py
```

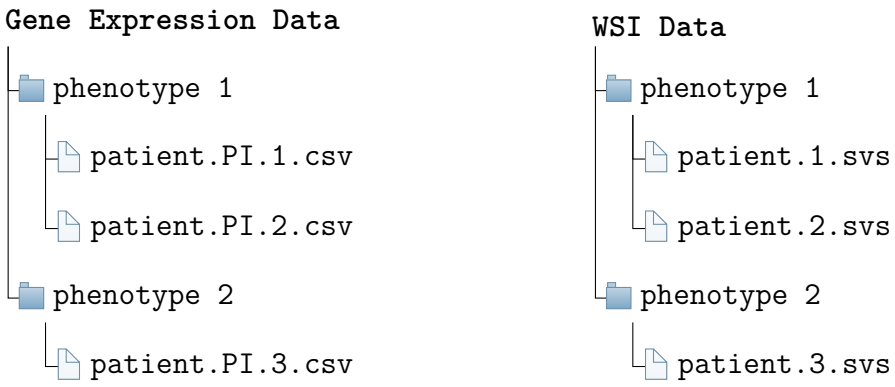


Fig. B.2 The directory tree structure for DL using bimodalities. Each main folder within the directory is labelled with a specific phenotype, containing subfolders that house PI files in comma-separated value (*CSV*) format for RNA-Seq data and slide virtual slide (*SVS*) format for WSI data. This organised structure ensures that the framework can correctly process the data, allowing for seamless integration of both modalities in the DL model.