Exploring Topological Data Analysis in Gene Expression Data

Topology-Driven Biomarker Discovery and Clinical Outcome Prediction in Oncology

Ndivhuwo Nyase

Supervisor: Dr Lebohang Mashatola
Dr Stephanie Julia Muller
Dr Musalula Sinkala

Department of Statistical Sciences
University of Cape Town

This minor dissertation is submitted for the degree of
Master of Science

June 2025

This thesis is dedicated to my late father, Dr. Muvhulawa Simon Nyase.

Declaration

I, Ndivhuwo Nyase hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Ndivhuwo Nyase
June 2025

Acknowledgements

This MSc has been intellectually demanding and, at times, deeply challenging, but also profoundly rewarding. It would not have been possible without the unwavering support, guidance, and encouragement of many, for which I am immensely grateful. Above all, I thank God for giving me the strength, wisdom, and perseverance to navigate this journey.

I express my deep gratitude and appreciation to my mentor and supervisor, Dr. Lebohang Mashatola. I fondly recall the first time he introduced me to Topological Data Analysis, a moment filled with both curiosity and uncertainty. Now, having co-developed WGTDA with him, I fully appreciate your invaluable guidance and mentorship. Your expertise has shaped this thesis and significantly contributed to my personal and intellectual growth.

I extend my heartfelt gratitude to my supervisor and technical lead, Dr. Stephanie Muller, whose leadership and guidance have been instrumental in advancing this work. I sincerely thank my supervisor, Dr. Musulala Sinkala, for your support and insightful conversations. Your guidance and perspective have been invaluable in enriching my research.

I sincerely thank IBM Research Africa and my colleagues for their invaluable support, insight, and encouragement. Their guidance has significantly influenced both my research, personal and professional development. Their contributions have been a source of inspiration, and I truly appreciate the impact they have had on my journey.

To my friends, you are too many to mention, but if you've ever come to my home and been welcomed by my family, you know who you are. Your support, laughter, and companionship have meant the world to me, and I'm truly grateful to have you in my life. To my girlfriend, Lesedi Bogopa, thank you for your continued support and inspiration. To my cousin, Thami Ntshudisane, I'm forever grateful to have you as family.

To my sister and my mother, having you in my life is a blessing beyond words. Your love, encouragement, and unwavering belief in me have been my strength and motivation. I honestly don't know what I'd do without you.

Finally, I want to thank my father, Dr. Muvuhulawa Simon Nyase, for empowering me, listening to me, trusting me, and leading me. I am truly blessed to have you not only as a father but also as a best friend and a role model. Your wisdom, strength, and kindness continue to inspire me and guide me every day. Though you are no longer here, your presence is felt in everything I do, and I carry your lessons with me. I hope this achievement honors the legacy you left behind, and I will continue to honor you and make you proud.

Abstract

This thesis is grounded in the fundamental observation that biological data has shape and this shape matters. Beneath the high-dimensional, often noisy landscape of gene expression profiles lie hidden topological structures (connected components, loops and voids) that capture the complex relationships driving cancer development and progression. By embracing this perspective, we position Topological Data Analysis (TDA) and persistent homology at the core of a novel analytical framework designed to tackle two key challenges in cancer research: clinical outcome prediction and biomarker discovery.

In this study, we employ Weighted Gene Topological Data Analysis (WGTDA) to extract topological features from gene expression data, which serve as prognostic biomarkers for cancer classification, staging, and treatment response. Moreover, by integrating these topological features with machine learning models we aim to enhance the predictive accuracy for clinical outcomes.

For clinical outcome prediction, we transformed gene expression profiles into topological fingerprints using multiple co-expression measures—namely, Pearson Correlation, Distance Correlation, and Weighted Topological Overlap (wTO) computed with both Pearson and Distance-based adjacencies. These topological features were analyzed using Random Forests. In parallel, we compared the predictive performance of traditional machine learning models (SVM, Gradient Boosting Decision Trees, Random Forest, and Neural Networks) trained on raw gene expression data against models incorporating the topological fingerprints. This comparative analysis was conducted across three classification tasks: cancer type (using TCGA-SARC, TCGA-PCPG, and TCGA-ESCA datasets), cancer staging (using TCGA-HNSC for stages I–IV), and treatment response (responders vs. non-responders).

For biomarker identification, the same three tasks were applied using the best-performing co-expression measure to generate a global topological representation of the patient population. This provided a disease-level view, highlighting shared homological patterns to facilitate biomarker discovery. Additionally, a dedicated visualization tool has been developed to aid in interpreting these topological signatures and identifying critical biomarkers. The tool is available at https://nnyase.github.io/MSc-Thesis/

WGTDA significantly enhanced phenotype prediction tasks by overcoming common pitfalls of traditional ML models in RNA-Seq data, such as overfitting and poor handling of class imbalance.

TDA-derived features improved generalizability of ML models in tasks such as cancer staging and treatment response prediction. Our findings strongly support the integration of TDA into clinical outcome prediction, demonstrating its value in capturing nuanced patterns that allow ML methods to learn more effectively.

Moreover, WGTDA remarkably identified key gene signatures for cancer type, staging, and treatment response without relying on pre-existing biological assumptions—yielding biomarkers that are strongly supported by the existing literature. These results underscore the method's reliability and potential clinical utility in precision oncology.

Table of contents

List of Figures

List of Tables

Nomenclature

5-FU	5-fluorouracil
AI	Artificial Intelligence
β	Betti Number
BH	Benjamini-Hochberg
BRCA	Breast Cancer
CAD	Computer-aided diagnosis
cDNA	Complementary Deoxyribonucleic Acid
CNN	Convolutional Neural Networks
COAD	Colorectal Adenocarcinoma
CPTAC	Clinical Proteomic Tumor Analysis Consortium
DEGs	Differentially Expressed Genes
DNA	Deoxyribonucleic acid
EDA	Exploratory Data Analysis
EFB	Exclusive Feature Bundling
EHR	Electronic Health Records
EMT	epithelial-mesenchymal transition
ER	Estrogen Receptor
ESCA	Esophageal Carcinoma
ESCC	Esophageal squamous cell carcinoma

FC Fold Change

FPKM Fragments Per Kilobase Million

GBDT Gradient Boosting Decision Tree

GDC Genomic Data Commons

GEO Gene Expression Omnibus

GLM Generalized Linear Model

GOSS Gradient-Based One-Side Sampling

GRN gene regulatory networks

HCC hepatocellular carcinoma

IBM International Business Machines

IHC Immunohistochemical

IPA Ingenuity Pathway Analysis

LightGBM Light Gradient Boosting Machines

LTR likelihood Ratio Test

LUAD Lung Adenocarcinoma

ML Machine Learning

MSE Mean Squared Error

NB Negative Binomial

NCBI National Center for Biotechnology Information

NGS Next Generation Sequencing

NLP Natural Language Processing

NN Neural Network

OS Osteosarcoma

PCA Principal Component Analysis

PCPG Paragangliomas/Pheochromocytomas

$\ltimes$ Set of natural numbers

$\mathbb{R}$ Set of real numbers

RA Research Aim

RB Retinoblastoma

RF Random Forest

RNA-Seq Ribonucleic Acid-Sequencing

RNA Ribonucleic acid

SARC Sarcoma

STRING Search Tool for the Retrieval of Interacting Genes/Proteins

SVM Support Vector Machines

t-SNE t-Distributed Stochastic Neighbor Embedding

TCGA The Cancer Genome Atlas

TDA Topological Data Analysis

TNM Tumour, Node, Metastasis

TOM Topological Overlapping Measures

UMAP Uniform Manifold Approximation and Projection

VEGF vascular endothelial growth factor

VR Vietoris-Rips

WGCNA Weighted Gene Co-expression Network Analysis

WGTDA Weighted Gene Topological Data Analysis

WHO World Health Organisation

wTO Weighted Topological Overlap

Chapter 1

Introduction

1.1 Background

I encourage you to take a moment and look around. The physical world is filled with shapes — from the contours of a mountain to the structural design of a bridge, and the intricate folding structure of a protein molecule, these shapes and structures are fundamental to how we understand and interact with the world around us [1]. This principle extends beyond the physical world as it permeates into the domains of science where the study of shapes - topology - becomes a powerful tool in analyzing complex systems and data. Topology is the mathematical study of shapes and spatial properties that remain invariant under continuous deformations [2]. To many, it may seem like a fascinating but abstract and impractical field of study. Perhaps topology was only a brief module in a dense mathematics course in university. However, upon closer examination, topology reveals itself as both ubiquitous and deeply embedded in nature, science and data. Its principles are essential for unraveling the hidden patterns, structures and relationships that govern complex systems we seek to understand. In an era dominated by big data and interconnected relationships, topology stands not as an obscure mathematical curiosity, but as a transformative framework with profound implications for scientific discovery.

This is where Topological Data Analysis (TDA) comes into play. TDA leverages the concepts of algebraic topology to uncover and analyze intricate structures and features embedded in complex and high-dimensional data [3]. At the heart of TDA lies persistent homology, a technique that tracks the emergence and closure of topological features across different scales. Conventional bioinformatics, statistical, and machine learning techniques often rely on linear correlations, network analysis, or iterative optimization processes like backpropagation. While effective in many scenarios, these methods often encounter limitations when confronted with high dimensional data, noise and the interpretability of results [4–6]. On the contrary, TDA focuses on capturing both local and global topological features, thus extracting structural properties embedded in data that conventional techniques may overlook. By analyzing the topology of the data, TDA identifies key homological and

geometric features, such as cycles, loops, voids, tunnels and higher dimensional surfaces, offering a different perspective to data [7, 8]. These topological insights are particularly useful in the context of omics data, where understanding the interactions between molecular components is key to uncovering important biological processes.

Omics studies provide a means to decipher the intricate interactions within cellular and molecular systems that drive diseases such as cancer. As one of the most prevalent non-communicable diseases worldwide, cancer remains a critical global health concern [9]. As such, omics studies are crucial for identifying novel biomarkers and accurately predicting clinical outcomes, both of which are important for advancing drug discovery and bridging the gap between cell biology research and clinical applications [10].

However, identifying biomarkers and predicting clinical outcomes poses a formidable challenge due to the high-dimensional nature of omics data [11, 12]. Additionally, researchers must discern meaningful signals from noise and validate biomarkers across diverse populations and conditions. Moreover, analytical techniques are required to detect subtle yet clinically relevant patterns within the data [13]. These challenges are particularly evident in gene expression studies, where the sheer volume and complexity of the data necessitate innovative analytical frameworks [11, 12]. This is where TDA emerges as a transformative approach. By characterizing topological spaces and structures in omics data, TDA can help interpret biological processes, potentially leading to earlier and more accurate disease diagnoses, a better understanding of disease mechanisms, and overall improvements in patient care.

The primary goal of this thesis is to leverage TDA to advance biomarker discovery and enhance clinical outcome predictions in cancer research through the analysis of gene expression data. As such, this thesis focuses on addressing two primary research objectives. First, we explore the application of TDA for biomarker discovery through the use of Weighted Gene Topological Data Analysis (WGTDA), which I co-developed with IBM Research [14, 15]. WGTDA serves as both a data mining methodology and an open-source Python package, designed to uncover key topological features within gene expression data that can potentially serve as prognostic biomarkers for complex diseases[1][2]. Additionally, we are developing an accompanying visualization tool to facilitate the analysis and interpretation of these topological features, providing an intuitive means to explore gene-phenotype relationships. Once identified, these topological interactions will undergo validation using external literature to assess their prognostic value and clinical relevance. Importantly, we focus on uncovering biomarkers associated with distinct cancer types, cancer stage, and treatment response, laying the foundation for downstream predictive modeling.

[1] For access to the WGTDA tool on GitHub: github.com/IBM/WGTDA
[2] For the original paper on WGTDA, see: arxiv.org/abs/2402.08807

Building on these insights, the second research objective focuses on extending the application of TDA by integrating topological features with machine learning models to predict key clinical outcomes. These outcomes mirror the three tasks addressed in the first objective — cancer type, cancer stage, and treatment response, with each becoming progressively more difficult to determine. While the cancer type is relatively easier to predict, cancer stage requires a deeper understanding of tumor progression, and treatment response presents the most significant challenge due to the complex and often unpredictable biological factors involved. In addressing this objective, we systematically compare the performance of machine learning models both independently and in combination with topological features. This comparative analysis evaluates how the inclusion of topological descriptors impacts predictive accuracy across these diverse clinical outcomes.

Therefore, this thesis demonstrates how TDA can be effectively applied to both biomarker discovery and clinical outcome prediction. Through the WGTDA method, we approach biomarker discovery from a disease-level by converting the gene expression profiles of all patients into a single topological space. This global disease view allows us to identify shared patterns and structures across the entire patient population. In contrast, for predicting clinical outcomes, each patient is assigned a unique topological fingerprint, derived from their individual gene expression profile. This patient-specific representation captures variations and nuances essential for personalized predictions of cancer type, cancer stage, and treatment response. Together, these complementary approaches showcase the versatility and power of TDA and WGTDA, bridging population-level biomarker discovery with individualized clinical predictions, ultimately contributing to more precise diagnostics and identification of better biomarkers.

1.2 Research Objectives

1.2.1 General Aim

The overarching aim of this thesis is to leverage TDA to advance biomarker discovery and improve clinical outcome predictions in cancer research through the analysis of gene expression data. This approach seeks to uncover prognostic biomarkers and improve the predictive power of machine learning models in clinical applications.

1.2.2 Specific Aims

This study is structured around two interrelated areas: Biomarker Discovery and Clinical Outcomes. Each area addresses key challenges in applying TDA to gene expression and cancer research and, together, they form a comprehensive framework. The specific aims are as follows:

- **Biomarker Discovery**

    1. *Identify prognostic biomarkers in gene expression data using WGTDA for cancer type, cancer stage and treatment response.*
    2. *Develop a visualization framework to represent topological features in gene expression as a complex network.*

- **Clinical Outcomes**

    1. *Evaluate the impact of integrating topological descriptors on the predictive performance of machine learning models for clinical outcomes.*
    2. *Compare different co-expression measures in constructing the simplicial complex and assess their influence on model performance.*

The following section provides a detailed exploration of each aim, outlining their significance, and expected contributions to biomarker discovery and clinical outcome prediction.

1.2.3 Biomarker Discovery

1.2.3.1 *Identify prognostic biomarkers in gene expression data using WGTDA for cancer type, cancer stage and treatment response.*

A key objective in biomarker discovery is determining whether an analytical method can reliably identify clinically relevant markers across diverse phenotypes [16]. In this study, the phenotypes under investigation are cancer type, cancer stage, and treatment response and they vary in complexity. By employing WGTDA, we aim to extract topological features from gene expression data that serve as robust biomarkers for these phenotypes. To establish the practical utility of WGTDA, we will validate the identified biomarkers against published literature, assessing both their clinical significance and generalizability. We hypothesize that WGTDA will uncover gene signatures associated with the proposed phenotypes, establishing it as a powerful framework for biomarker discovery across diverse clinical contexts.

1.2.3.2 *Develop a visualization framework to represent topological features in gene expression as a complex network.*

In tandem with identifying biomarkers, visualizing gene expression networks is critical for deciphering the complex interactions and relationships between genes, proteins, and other molecular entities. These visual representations are a powerful means of interpreting biological systems, enabling researchers to map gene-gene interactions, uncover molecular pathways, and pinpoint potential biomarkers that are pivotal in disease phenotypes or therapeutic responses [17, 18].

Tools such as Cytoscape [19], Ingenuity Pathway Analysis (IPA) [20] and STRING database [21] have become important for visualizing these networks. In this thesis, we aim to utilize the topological features derived from WGTDA to construct and visualize gene networks similar to the tools mentioned above. We hypothesize that such topological gene networks will reveal critical genes and key patterns, thereby facilitating a deeper exploration of gene-phenotype relationships and broadening our capacity for effective biomarker discovery using TDA.

1.2.4 Clinical Outcomes

1.2.4.1 *Evaluate the impact of integrating topological descriptors on the predictive performance of machine learning models for clinical outcomes.*

Predicting clinical outcomes with machine learning is challenging due to the high dimensionality and complexity of gene expression data, which can lead to overfitting and reduced generalizability [11, 12]. To address these issues, we propose incorporating topological descriptors derived from TDA as features in random forest models. By integrating these descriptors, we aim to enhance model robustness and improve predictive accuracy for key clinical outcomes, including cancer type, cancer stage, and treatment response.

To test this hypothesis, we conduct a comparative analysis of machine learning models with and without TDA-derived topological features. The evaluation includes the following models:

- Support Vector Machines (SVM)
- Light Gradient Boosting Machines (LightGBM)
- Neural Networks (NN)
- Random Forest (RF)

This analysis will assess whether integrating topological descriptors improves traditional machine learning models for clinical prediction tasks and measure the extent of their impact.

1.2.4.2 *Compare different co-expression measures in constructing the simplicial complex and assess their influence on model performance.*

One critical challenge in TDA is selecting the most appropriate measure to construct the simplicial complex. A simplicial complex is a fundamental topological structure that encapsulates both geometric and relational properties of the data. While we formally define the simplicial complex in the Theoretical Framework (Chapter 2), it is essential to note that the choice of metric significantly affects the resulting topological features. Different measures may emphasize different patterns within the data, making the selection process non-trivial and application-specific. Despite its importance, there is a notable gap in the literature regarding systematic comparisons of these measures for

constructing informative simplicial complexes for downstream tasks [22].

To address this gap, we investigate the utility of four co-expression measures:

- Pearson's Correlation
- Distance Correlation
- Weighted Topological Overlap (wTO) using Pearson-based adjacency
- Weighted Topological Overlap (wTO) using Distance-based adjacency

Our goal is to evaluate how each measure influences the construction of topological descriptors and, consequently, the performance of the machine learning models in predicting clinical outcomes. By systematically comparing these approaches, we aim to identify the most informative co-expression measure for clinical prediction tasks.

We will then adopt the best-performing measure to generate global, population-level biomarkers using WGTDA, directly addressing research aim 1.2.3.1. This strategy ensures that the resulting biomarkers are maximally robust and biologically meaningful.

To our knowledge, this is the first work to apply TDA to different phenotype predictions while also integrating a visualization framework for biomarker discovery. This innovative approach highlights the versatility of TDA in genomics and personalized medicine, effectively bridging the gap between fundamental TDA research and clinical applications within a unified analytical framework.

1.3 Motivation

The rapid advancement of digital technologies, coupled with the decreasing costs of sequencing and data acquisition, has resulted in an explosion of biological data. This data explosion encompasses next-generation sequencing (NGS), enhanced medical imaging, and a growing emphasis on personalized medicine [23]. Simultaneously, the scientific community has embraced a culture of openness, fostering the development of open-source tools and the sharing of vast datasets in publicly accessible repositories. Prominent examples of these repositories include the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) [24] and Genomic Data Commons (GDC) [25]. This era of big biological data has placed a need for advanced analytical techniques capable of extracting meaningful insights from these vast datasets [6]. Although artificial intelligence (AI), machine learning (ML), and statistics have advanced our understanding of big biological data, effectively interrogating complex, high-dimensional datasets still remains challenging [11, 12]. In this context, TDA has emerged as a promising approach for exploring the shape and structure of

high-dimensional omics datasets [26].

With that being said, while TDA offers a powerful framework for uncovering hidden structures in data, it has not gained the same widespread recognition or level of adoption as other data-driven techniques, such as AI, ML and statistics. The irony is that the same intricate and rigorous theoretical framework that gives TDA its strength also renders it less accessible to researchers outside of pure mathematics. As a result, TDA has gained far less recognition compared to more user-friendly and widely adopted techniques, limiting its adoption and application in life sciences research [26].

Moreover, a natural extension of this is the recognition that biological data itself, much like the biological entities, has an inherent shape and structure that can reveal valuable insights. Researchers often study how the shapes of biomolecules, cells, tissues, and organisms arise from the effects of genetics, development, and environmental factors. Consequently, it is less common to consider that biological data also has shape and structure [27]. Just as biological forms can be analyzed and interpreted, so too can the structure of data be measured, explored and transformed to uncover underlying patterns and hidden relationships. Recognizing that data has shape, particularly in high-dimensional and noisy datasets, offers a new and powerful perspective that opens doors to novel discoveries in biology and medicine [23].

However, these challenges also present an opportunity to advance biological research by positioning TDA as a robust methodology for analyzing complex biological data, such as gene expression. In this thesis, we embrace this challenge by proposing that biological data is best understood through the lens of shapes, networks, interactions, and homology groups. As data scientists, we bear the responsibility to rigorously test and validate new methodologies across diverse applications to ensure their robustness and broad applicability. By doing so, we aim to bridge the gap between theoretical advancements and practical use, contributing to the wider adoption of TDA in life sciences and demonstrating its value in analyzing complex biological systems. Ultimately, this thesis is grounded in the belief that biological data has shape—and that shape matters.

1.4 Thesis Outline

- **Chapter 1: Introduction**

  This chapter establishes the foundation of the thesis by introducing topology and TDA. It outlines the research aims, presents the motivation behind the study, and highlights the significance of applying TDA in biomarker discovery and clinical outcome prediction.

- **Chapter 2: Theoretical Framework**

This chapter provides the theoretical underpinnings of the research, introducing key topological concepts such as simplicial complex, persistent homology, and *betti* numbers. These mathematical foundations are essential for understanding how TDA is applied to omics datasets.

- **Chapter 3: Literature Review**

  This chapter critically examines the current state of research in areas central to this thesis, including cancer biology, statistics, machine learning and bioinformatics. It explores key related works that inform this study while identifying existing gaps and opportunities that this research aims to address.

- **Chapter 4: Methods**

  This chapter details the research methodology for both biomarker discovery and clinical outcome predictions. It covers data acquisition, preprocessing and preselection steps, co-expression measures, and the corresponding TDA pipeline, evaluation metrics and the visualization framework built for WGTDA.

- **Chapter 5: Results**

  This chapter presents the experimental findings for both biomarker discovery and clinical outcome prediction. Exploratory Data Analysis (EDA), visualizations, statistical analyses, and comparative results are included to support the conclusions drawn.

- **Chapter 6: Discussion**

  This chapter provides an in-depth interpretation of the results, connecting them to the research objectives and situating them within the broader context of cancer research and bioinformatics. The implications, limitations, and potential future directions of the study are also discussed.

- **Chapter 7: Conclusion**

  This chapter summarizes the key findings of the thesis, emphasizing its contributions to TDA, biomarker discovery, and clinical outcome prediction. It concludes by highlighting the impact of the research and its potential for future applications in bioinformatics, oncology and TDA research.

Chapter 2

Theoretical Framework

In this chapter, we establish the theoretical foundation underpinning the methodologies employed in this thesis. We introduce essential concepts from topology and TDA that are crucial for uncovering the intrinsic structure of complex data. In particular, we examine key constructs such as network theory, simplicial complexes, persistent homology, and *betti* numbers, that serve as powerful tools for this purpose. This framework is vital for understanding the methods used to achieve the research objectives outlined in Chapter 1.

2.1 Topology

As in any branch of mathematics, determining which objects are equivalent is essential. In topology, this equivalence is explored using properties such as continuity, compactness, and connectedness [28, 29]. Topology studies the inherent features of geometric objects that remain unchanged under continuous deformations—such as stretching, twisting, crumpling, or bending, without concern for precise measurements or size [2, 28, 29]. A topological space is a collection of open sets equipped with a topology. This space provides a framework that enables us to define and analyze concepts like continuity, convergence, and connectedness independent of any notion of distance [28, 29]. Common examples of topological spaces include the Euclidean spaces ($\mathbb{R}^n$) and metric spaces, where the topology is induced by a distance function, or a measure. This discipline, often called "rubber-sheet geometry," focuses on an object's connectivity and overall structure rather than its exact size and measurement details.

These observations naturally lead us to ask: How can we classify topological spaces? One powerful approach is to examine their "holes" or voids. The presence or absence of holes provides crucial insight into the structure of a space and serves as an invariant under continuous deformations. This concept is central to algebraic topology, which offers robust methods for detecting, counting, and comparing these holes [28, 30].

A well-known joke in mathematics captures the essence of this approach: "Topologists can't tell the difference between a coffee mug and a donut!" This joke highlights the essence of topology. As illustrated in Figure 2.1, a coffee mug and a donut are topologically equivalent because each possesses a single hole (the mug's handle and the donut's central hole). Consequently, through continuous deformation a coffee mug can be transformed into a donut, and vice versa without any tearing or re-gluing.

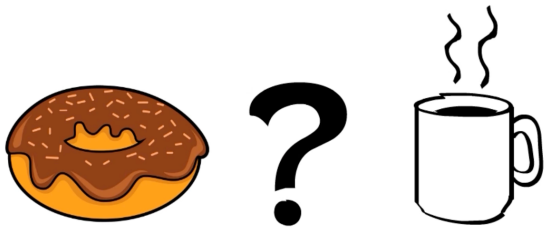


Fig. 2.1 A topologist's favorite joke

Having established the foundational concepts of topology and the role of holes in classifying spaces, we now turn to how these ideas can be applied to real-world data. TDA extends these principles to uncover complex structures and relationships within datasets, revealing hidden patterns while preserving their geometric and connectivity properties.

2.2 Topological Data Analysis (TDA)

At its core, TDA translates raw data into topological spaces, where we can analyze its shape and connectivity. Just as topology classifies spaces based on properties that persist under continuous deformations, TDA identifies stable topological features such as connected components, loops, holes and voids within data. These features provide a robust representation of the dataset's structure across multiple scales [31–33].

A key technique in TDA is persistent homology, which tracks how topological features evolve as we vary a scale parameter [32, 34]. Essentially, we construct a sequence of nested topological spaces using data, and analyze how features like holes appear and disappear. The persistence (lifespan) of these features is represented in a persistence diagram or persistence barcode, providing a concise summary of the dataset's topological structure [30].

In essence, TDA extends the principles of topology beyond theoretical spaces to real-world data, allowing us to classify and interpret complex structures through the lens of topological invariants. To effectively apply these ideas, we must first establish how data points relate to one another—leading us to network theory, which provides a foundational framework for representing complex relationships within datasets.

2.3 Network Theory

Networks are essential for representing complex systems by capturing the relationships between elements within a dataset. Formally, a network is defined as a graph $G(V,E)$ where V is a set of nodes or vertices ($v_i \in V$) and E is a set of edges or links ($e_i \in E$) that connect pairs of vertices. The connectivity of a network is often represented through an adjacency matrix $\mathbf{A} = a_{i,j}$, where $a_{i,j} \neq 0$, if there exists a connection between the pairs of vertices. By capturing interactions, dependencies, and associations within a system, networks become a powerful framework for analyzing complex datasets.

While networks are typically represented as graph, it can also be used to form a topological space. [35]. By analyzing the relationships encoded within a network's structure, we can extract meaningful topological features that reveal the underlying shape of the data.

2.3.1 Weighted Networks

An important extension of networks is the concept of a weighted network, denoted $G(V,E,W)$. In a weighted network each edge $e_{i,j} \in E$ connecting vertices v_i and v_j is assigned a weight $w_{i,j} \in W$ that represents the strength, frequency, or significance of the relationship between the connected nodes v_i and v_j. he weights provide additional insight into the connections between pairs of nodes, allowing for a more nuanced representation of the underlying data.

2.3.2 Network Filtration

To analyze how a network's topology changes given different scales, parameters or thresholds, we employ a procedure known as network filtration. This approach is relevant for weighted networks, where each edge's weight influences how and when that edge enters or exits the network [35, 36].

Network filtration is a process that systematically tracks geometric and network properties as the threshold values imposed on the network vary. In practice, this method involves constructing a series of networks by iteratively adding or removing edges below a specific threshold. For example, edges with weights lower than 0.2, or 0.8 are added or removed. By generating these threshold graphs, we can identify persistent network metrics - local and global properties of the network that remain stable across a range of threshold values. These persistent features can be plotted or tracked throughout the filtration process, providing valuable insights into the network's structure [35, 36]. An intuitive example of network filtration is analyzing social networks. By examining how connections persist across different threshold levels, we can gain insights into the strength and stability of social ties within a group.

This approach mirrors the concept of filtrations in topology, where one explores how a topological space evolves by continuously adding or removing elements. To better understand the structural

evolution of these spaces, we must first introduce their fundamental building blocks: simplices and simplicial complexes

2.4 Simplex

A simplex is a geometric construct that generalizes the notion of points, triangles or tetrahedrons to arbitrary dimensions. It serves as a foundational element in the construction of simplicial complexes, which are crucial for analyzing topological properties in data [29, 37].

Formally, a k-simplex is defined as the convex hull of $k+1$ affinely independent vertices in $\mathbb{R}^n$. The dimension of the simplex corresponds to the number of vertices minus one. Simplices can exist in infinite dimensions. As illustrated in Figure 2.2, we begin with a 0-simplex (point), and it evolves through higher dimensions including a 1-simplex, 2-simplex, and a 3-simplex. These figures demonstrate how simplices encode geometric structures across varying dimensions.

- A 0-dimensional simplex is a single point.
- A 1-dimensional simplex is a line segment connecting two points (an edge).
- A 2-dimensional simplex is a triangle, which includes its three edges and the interior region.
- A 3-dimensional simplex is a tetrahedron, the 3-dimensional analog of a triangle, consisting of four triangular faces and the enclosed volume.

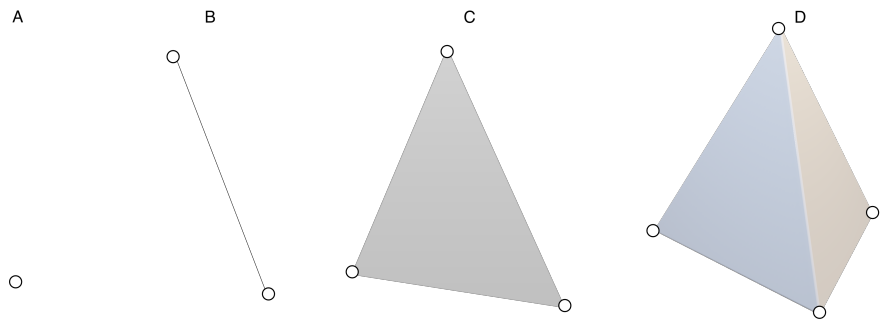


Fig. 2.2 Simplices of different dimensions. (A) a vertex (0-simplex), (B) an edge (1-simplex), (C) a triangle (2-simplex), and (D) a tetrahedron (3-simplex). We note that an edge has two vertices, a triangle has three edges, and a tetrahedron has four triangles as faces.

Suppose the points $u_0, u_1, \ldots, u_k$ are affinely independent, meaning the vectors $u_1 - u_0, u_2 - u_0, \ldots, u_k - u_0$ are linearly independent. The simplex determined by these points is the set of points C defined as:

$$C = \left\{ \theta_0 u_0 + \cdots + \theta_k u_k \;\middle|\; \sum_{i=0}^{k} \theta_i = 1 \text{ and } \theta_i \geq 0 \text{ for all } i = 0, \ldots, k \right\}. \tag{2.1}$$

The equation 2.1 represents the convex combination of the points $u_0, u_1, \ldots, u_k$, where the coefficients $\theta_0, \ldots, \theta_k$ are non-negative and sum to 1. The convex combination ensures that the resulting point lies within the simplex, rather than outside of it.

To elaborate on definition 2.1:

1. **Convex Hull:** The convex hull of a set of points is the smallest convex set that contains all the points.
2. **Affine Independence:** The points $u_0, u_1, \ldots, u_k$ are affinely independent, meaning the vectors $u_1 - u_0, u_2 - u_0, \ldots, u_k - u_0$ are linearly independent. This ensures that the points span a space of the correct dimension (e.g., three non-collinear points in $\mathbb{R}^2$ form a triangle).
3. **Convex Combination Condition:** The condition $\sum_{i=0}^{k} \theta_i = 1$ ensures that we are forming a point inside the convex hull of the given points. Each θ_i is a non-negative scalar that assigns a weight to each point u_i, and the sum-to-one condition guarantees that the point lies within the boundary of the simplex.

Simplices are the building blocks of simplicial complexes, which allow for the computation of homology and the extraction of topological features from data [34]. By combining simplices of different dimensions, a simplicial complex can reveal intricate structures and relationships within the data.

2.5 Simplicial Complex

A simplicial complex is a fundamental mathematical construct that extends the concept of a simplex by assembling simplices in a structured, combinatorial manner. It serves as a framework for representing geometric and topological spaces within data. Intuitively, a simplicial complex can be visualized as a structure built from points (0-simplices), edges (1-simplices), triangles (2-simplices), and their n-dimensional counterparts, all interconnected to form a coherent mathematical object [29, 30].

Formally, a simplicial complex X in $\mathbb{R}^n$ is defined as a finite collection of simplices that satisfies the following two conditions as outlined in Elements of Algebraic Topology by James Munkres [29]:

1. **Face Condition:** Every face of a simplex in X must be an element of X. This ensures that if a simplex is part of a complex, then all of its lower-dimensional faces (such as edges and vertices) are also included in the complex.

2. **Intersection Condition:** The intersection of two simplices in X is either empty or is the face of both simplices. This means that if two simplices share a common part, that part must be a face of each simplex involved. If they do not share a common part, it must be empty.

These conditions ensure that the simplices are consistently connected, avoiding ambiguities or overlaps that violate the integrity of the complex.

Figure 2.3 illustrates a valid simplicial complex with a filtration at $K = 0.06$. The structure includes points, line segments, triangles, and tetrahedrons, all of which satisfy the two key conditions for a simplicial complex: the face condition, where every face of a simplex is included in the complex, and the intersection condition, where any two simplices intersect only at a shared face or not at all. In contrast, Figure 2.4 depicts a geometric configuration that fails to meet both the face and intersection conditions. Specifically, one simplex is missing a face, and two simplices intersect improperly in a non-face region. These violations render the configuration invalid as a simplicial complex.

Several types of simplicial complexes exist, each with unique properties and computational requirements, including the Vietoris-Rips (VR) complex [38], Čech complex [39], Alpha complex [40], and Witness complex [41]. Among these, the Vietoris-Rips (VR) complex is particularly well-suited for analyzing high-dimensional data due to its computational efficiency in calculating persistent homology through pairwise connections [42, 43].

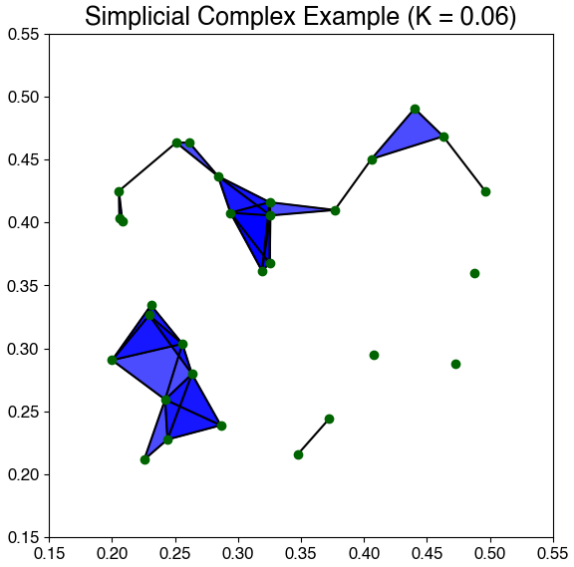


Fig. 2.3 Example of a simplicial complex with K or $\varepsilon = 0.06$. The structure includes points (0-simplices), edges (1-simplices), and filled triangles (2-simplices) that represent complete pairwise connections. This structure satisfies the conditions of a simplicial complex: (1) Every face of a simplex (e.g., edges and vertices of triangles) is included in the complex, and (2) the intersection of any two simplices is either empty or a shared face.

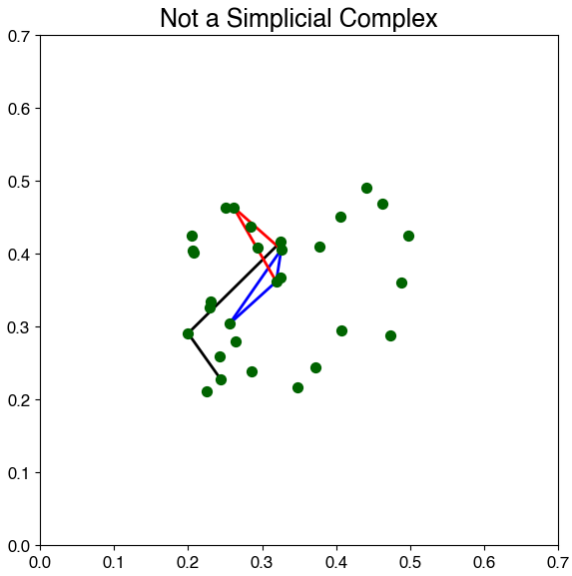


Fig. 2.4 Example of a structure that is not a simplicial complex. This structure violates two key conditions of a simplicial complex: (1) The face condition is violated as one triangle is missing an edge and (2) the intersection condition is violated as two triangles improperly share an edge without fully sharing a face.

2.5.1 Simplicial Filtration

To analyze how the topology of a simplicial complex evolves across different scales or thresholds, we use a method known as simplicial filtration. This process is analogous to network filtration (Section 2.3.2) and allows us to study the progressive evolution of the complex as additional simplices are incrementally added and connected. Mathematically, a filtration is represented as a sequence of nested simplicial complexes:

$$X = K_0 \hookrightarrow K_1 \hookrightarrow \ldots \hookrightarrow K_n \tag{2.2}$$

Here, K_0 represents the initial set of simplices, K_n represents the final complex after all simplices have been connected. The parameter K or ε serves as the filtration parameter, which determines when a new simplex is introduced. As K or ε increases, additional simplices are incorporated into the complex, and new topological features (e.g., connected components, loops, voids) emerge. These features can persist across multiple filtration scales, highlighting robust structures in the data, or they may disappear quickly, representing transient noise.

Figure 2.5 visually demonstrates the evolution of a VR complex as the filtration parameter K increases. The progression of VR complex showcases how increasing K gradually reveals more complex topological features. For instance:

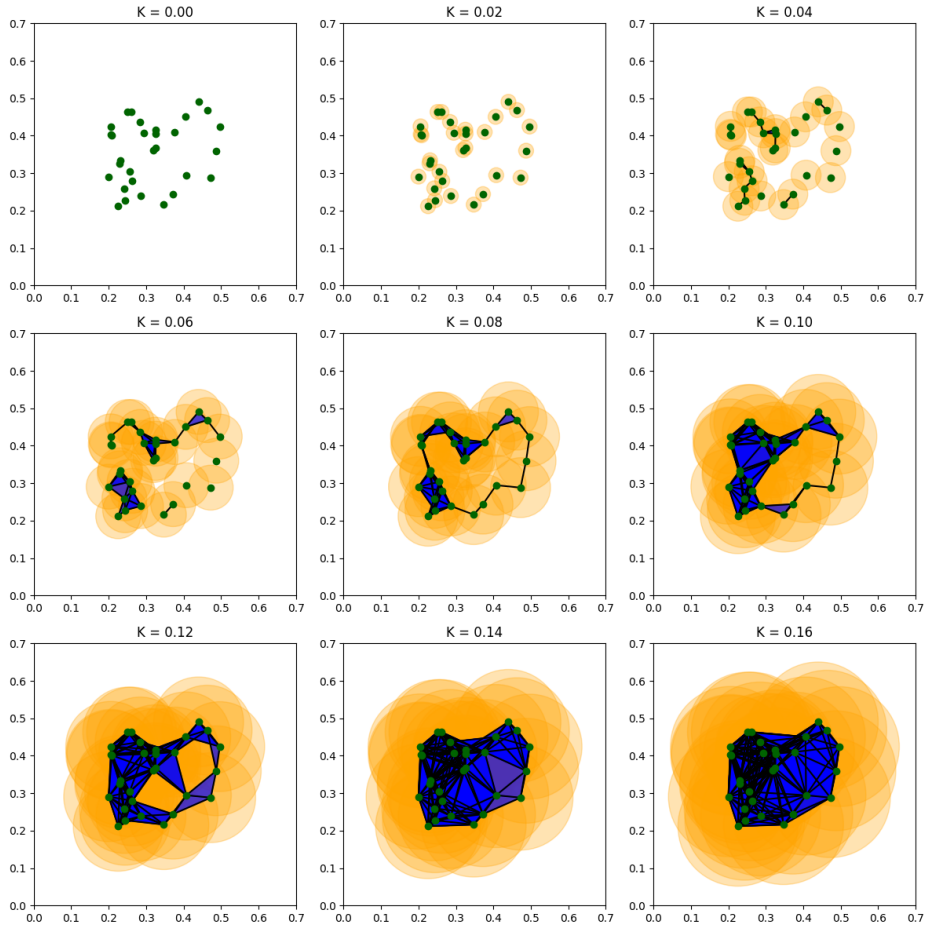


Fig. 2.5 Evolution of the VR complex. Each subplot represents a different threshold K or ε incrementing by 0.02, with circles illustrating the neighborhood radius, edges connecting points within K and filled triangles (2-simplices) highlighting complete pairwise connections. The progression showcases how increasing K gradually reveals higher-dimensional simplicial structures and more complex topological features.

1. At smaller thresholds (K), the complex consists mostly of disconnected points.
2. As K increases, these disconnected points become connected components. This is shown by the edges between points that begin to form.
3. At higher thresholds, loops, filled triangles (2-simplices) and tetrahedrons (3-simplices) emerge, marking the transition to higher-dimensional simplicial structures and more complex topological features.

2.6 Vietoris-Rips (VR) complex

A VR complex is a specific type of simplicial complex that is better-suited for high dimensional data analysis due to its computational efficiency [23, 43]. It is constructed by a set of points in $\mathbb{E}^n$ and ε, which represents a proximity threshold. For a given ε, an edge is drawn between every pair of points that are within a distance ε of each other.

Mathematically, for a set of points X, the VR complex $VR_\varepsilon(X)$ at scale ε is defined as follows:

- a simplex with vertices $\{x_0, x_1, ..., x_k\}$ is in $VR_\varepsilon(X)$ if and only if the pairwise distance between each pair of points x_i and x_j is less than 2ε.

The VR complex offers significant computational advantages over alternatives like the Čech complex. Unlike the Čech complex, which requires explicit calculation of higher-dimensional overlaps, the VR complex uses pairwise distances to infer the presence of higher-dimensional simplices, reducing computational overhead [23, 43].

2.7 Chains and k-Chain Groups C_k

We may ask: how do we translate topological spaces into algebraic objects for easier manipulation? This is achieved using chains and chain groups, which allow us to convert geometric and topological problems into algebraic ones. By operating on these chains, we can apply tools such as addition and subtraction of chains to gain deeper insights into the shape and connectivity of a space.

A chain is a formal linear combination of simplices of a certain dimension within a topological space, where the coefficients are elements from a specified coefficient group [28]. Intuitively, a chain represents a "sum" of geometric objects such as points, line segments, triangles, and other higher-dimensional simplices. A typical k-chain $c \in C_k$ is given by:

$$c = \sum_{i=0} a_i \sigma_i \tag{2.3}$$

where $a_i \in \mathbb{Z}$ and σ_i are the k-dimensional simplices.

Moreover, the chain group C_k is the collection of all possible k-chains.

$$C_k = \left\{ \sum_i a_i \sigma_i \,\middle|\, a_i \in \mathbb{Z},\ \sigma_i \text{ are } k\text{-simplices} \right\} \tag{2.4}$$

This group is free abelian, meaning it has a basis consisting of the k-dimensional simplices and every element in the k-chain can be expressed as a linear combination of the basis elements.

2.7.1 Boundary Maps and Boundary Homomorphism ∂_k

To analyze the topological features within a simplicial complex, we introduce the concept of a boundary of a chain and the associated boundary operators. Boundary maps and boundary homomorphisms are essential tools in algebraic topology for analyzing the structure of topological spaces. They track how higher-dimensional shapes are connected to their lower-dimensional boundaries [29].

For example, the boundary of a triangle (2-simplex) consists of its three edges (1-simplices), this is demonstrated in Figure 2.6 A. More generally, a boundary map of k is a linear function that sends a k-dimensional chain to its $(k-1)$-dimensional boundary.

$$\partial_k : C_k \to C_{k-1} \tag{2.5}$$

Importantly, the boundary map satisfies $\partial \circ \partial = 0$, meaning that the boundary of a boundary is always zero or empty (see Figure 2.6 B) [28, 29].

This property leads to the definition of *cycles* (chains with no boundary) and *boundaries* (chains that are the boundaries of higher-dimensional simplices) [1]. Homology groups are then computed by comparing cycles and boundaries, revealing key features such as holes and voids in the chain group.

2.8 Homology and Homology groups $H_n(X)$

Returning to the concept of holes or voids in a topological space, homology provides a bridge between geometry and algebra by captures the information about topological features within a space. Homology groups quantify the number of independent cycles in each dimension, allowing us to classify higher-dimensional analogs of loops and voids.

To compute homology, we begin by identifying chains within a simplicial complex. The boundary operator ∂_k introduced earlier maps a $k-chain$ to its boundary a $k-1$-chain. Formally, the k^{th}-homology group $H_n(X)$ is defined as the quotient of the space of k-dimensional cycles by the space of boundaries of $(k+1)$-dimensional simplices [28, 44, 45].

[1] Cycles and loops are terms used interchangeably throughout this thesis, both referring to closed chains or β_1 topological feature.

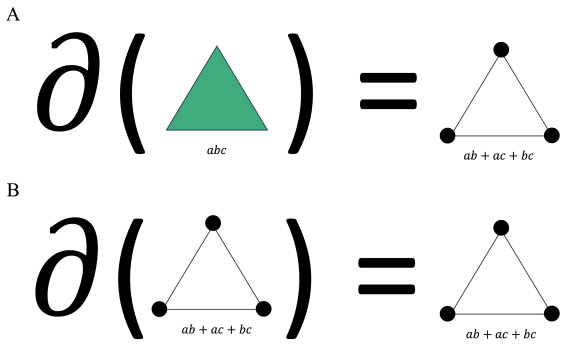


Fig. 2.6 Boundary operator of a triangle. In illustration A, the boundary map of a 2-simplex (triangle) is depicted, where the boundary of the triangle is the sum of its edges. Illustration B demonstrates that applying the boundary operator twice results in zero, confirming the principle that the boundary of a boundary is always zero.

$$H_n(X) = \frac{\ker(\partial_k)}{\text{img}(\partial_{k+1})} \tag{2.6}$$

These homology groups summarize the independent topological features at each dimension, providing a powerful way to analyze complex structures. For example, H_0 counts the connected components, H_1 identifies loops, and H_2 captures voids. By computing homology groups, we gain insight into the intrinsic shape of the data represented by the simplicial complex.

2.9 Betti Numbers β_k

Betti numbers are used to distinguish topological spaces based on the connectivity of n-dimensional simplicial complexes. To compute the *betti* numbers of a topological space, the space is first triangulated into a simplicial complex. This complex is then represented using a boundary matrices, and a matrix reduction process is applied to these matrices to determine the *betti* numbers. Specifically, the k-th *betti* number β_k of a simplicial complex K is defined as the rank of the k-th homology group $H_k(K)$, which quantifies the number of independent k-dimensional topological features of K [46]. This can be expressed as:

$$\beta_k = rank(H_k(X)) \tag{2.7}$$

Where *rank* is the number of independent generators of the k-th homology group ($H_k(X)$) representing the distinct k-dimensional topological features in the simplicial complex.

- β_0 represents the number of connected components (0-dimensional holes). In persistent homology they reveal isolated components that persist on different filtrations.
- β_1 represents the number of two-dimensional loops or holes. Persistent homology captures the birth and death of these loops through different filtrations
- β_2 indicates the number of three-dimensional voids, cavities or empty regions in the space. Again, persistent homology captures the birth and death of the cavities at various scales.

As depicted in Figure 2.7, *betti* numbers provide a summary of the topological features present in a topological space. For example, a point cloud consists of multiple disconnected points, which results in $\beta_0 = 9$, indicating nine connected components, with no higher-dimensional holes, so β_1 and $\beta_2 = 0$. A circle S^1 has a single connected component, represented by $\beta_0 = 1$, and a single 1-dimensional loop, giving $\beta_1 = 1$, while there are no 2-dimensional holes, hence $\beta_2 = 0$. In the case of a sphere S^2, we observe that it consists of one connected component, leading to $\beta_0 = 1$, and it encloses a 2-dimensional void, reflected by $\beta_2 = 1$, with no 1-dimensional loops, hence $\beta_1 = 0$. Finally, the torus $T^2 = S^1 \times S^1$ exhibits more complex topological orientation, with one connected component ($\beta_0 = 1$), two independent 1-dimensional loops ($\beta_1 = 2$), and a single 2-dimensional void ($\beta_2 = 1$).

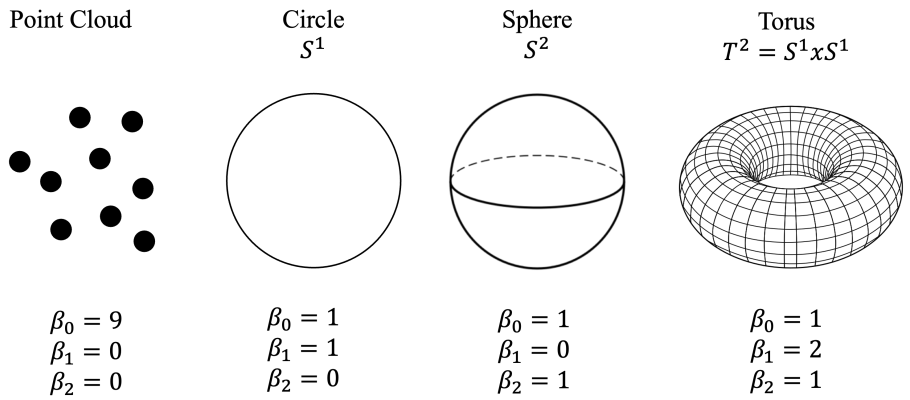


Fig. 2.7 In the illustration above, a point cloud has 9 connected components but no higher-dimensional holes, a circle S^1 has one connected component and one loop, a sphere S^2 has one connected component and one void, and a torus T^2 has one connected component, two loops, and one void.

Computing *betti* numbers involves algebraic techniques, primarily centered around reducing boundary matrices derived from a simplicial complex. Various algorithms have been developed for this purpose, including methods based on combinatorial laplacians [47], Smith Normal Form, and emerging approaches leveraging quantum algorithms [48, 49].

2.10 Persistent Homology

While homology and *betti* numbers provide a static view of topological features at a fixed scale, real-world data often requires an examination of how these features evolve across multiple scales. Persistent homology extends the homological framework by tracking the birth and death of topological features as a parameter ε varies, allowing for a multi-scale analysis of topological structures.

Persistent homology was introduced by Edelsbrunner, Letscher, and Zomorodian [44] as a method to compute homological features that persist over a range of scales, rather than features present at one single scale. The key idea is to build a filtration of simplicial complexes which are parameterized by ε. As ε increases, more simplices are added, and we track the birth and death of topological features. Essentially, the homology of the complex changes as ε increases; new connected components can appear, existing components can merge, loops and cavities can appear or be filled. The persistence or duration of these features across ε provides robust topological signatures that are stable under perturbations [44–46]. Thus, allowing one to analyze how topological features such as connected components (β_0), loops (β_1) and voids (β_2) appear and disappear across the filtration.

Persistent homology information is often encoded in a persistent diagram. A persistent diagram is a graphical representation of the birth and death intervals representing ε for a particular dataset. Essentially, the persistent diagram tracks the persistence of features by recording birth (b) and death scales (d) and can be defined as: $(b,d)\ |b,d \in R, 0 \leq b \leq d < \infty$. Each point (b,d) represents the birth and death scales of a topological feature. Features that persist over a large range (long lifespan $d-b$) are often considered significant, while those with short lifespans may be attributed to noise.

The process is illustrated in Figure 2.8, where a point cloud of five points arranged in a circular pattern undergoes persistent homology:

1. Initially in Figure 2.8A, only the individual points (β_0) are visible.
2. As ε increases, edges form, connecting nearby points.
3. At a critical threshold, the edges close to form a loop (β_1), capturing the circular structure.
4. The persistent homology evolution is summarized in the persistence diagram (Figure 2.8B), where each point represents the birth and death of a topological feature.

In persistent diagram depicted in 2.8B, the five black points correspond to β_0 representing the connected components that are born early and quickly die as they merge. The blue point, which persists much longer, corresponds to β_1 capturing the loop in the circular structure. Features with long persistence (greater difference between birth and death), such as the β_1 point, indicate significant structures, while those with short persistence, like the five individual black β_0 points, are often

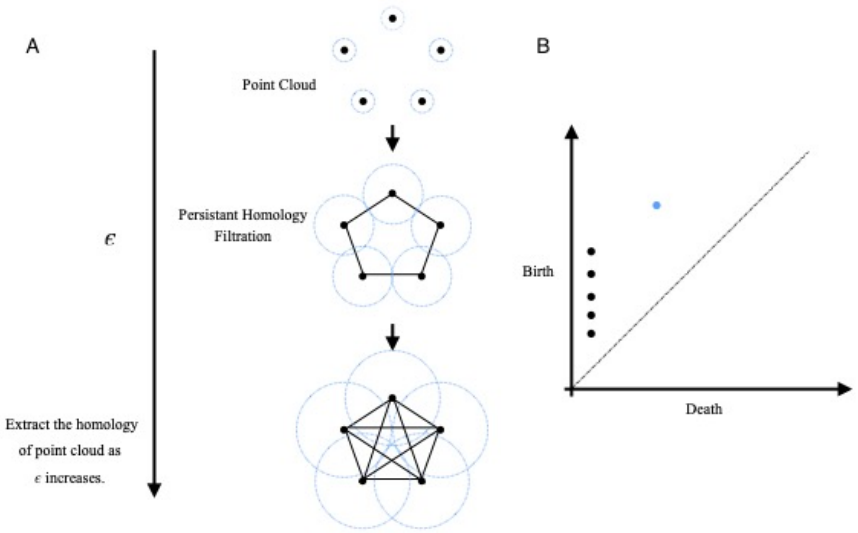


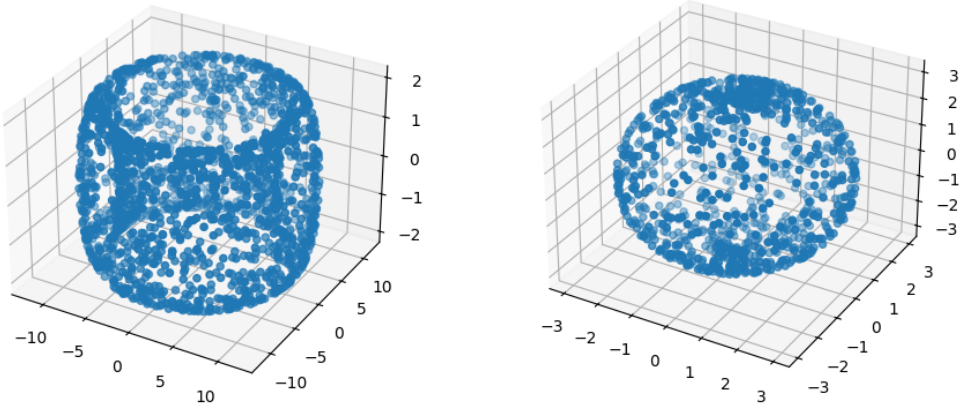
Fig. 2.8 Illustrates the process of persistent homology. The figure (A) shows a point cloud as a simplicial complex. Disks centered around each simplices grow iteratively as ε increases and when the disks intersect, and edge is formed. Connected points are β_0 topological features, and loops or holes are represented as β_1 features. The graph (B) is the persistence diagram which summarizes the birth and death coordinates each topological feature extracted in the point cloud, colored by β number; where black is β_0 and blue is β_1

attributed to noise.

To further understand persistent homology, we consider its application to two different point clouds, a torus and a sphere, shown in Figure 2.9. These point clouds represent samples taken from these shapes, and their persistent diagrams are provided in Figure 2.10. The key differences between the persistent homology of the torus and the sphere are evident in their Betti numbers and persistence diagrams:

1. ***Betti-1 (β_1)***: The torus (Figure 2.9a) is characterized by a prominent β_1 feature (orange point) that persists over a long range, representing a robust loop that encircles the torus longitudinally. This feature distinguishes the torus's topology by capturing its circular structure. In contrast, while the sphere (Figure 2.9b) has many β_1 features, it lacks any significant β_1 features, as they have small lifespans reflecting its absence of non-trivial loops.

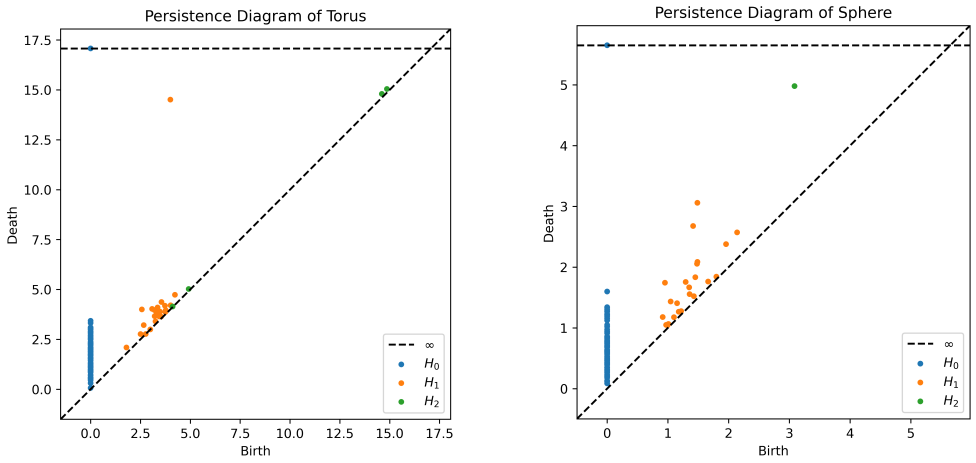
2. ***Betti-2 (β_1)***: Moreover, the torus displays multiple β_2 features (green points), corresponding to its hollow interior and complex 3D structure. On the other hand, the sphere has a single dominant β_2 feature, representing its central cavity or void. This simpler topological signature demonstrates the fundamental differences between the torus and the sphere in terms of their geometric and topological properties.



(a) A point cloud sampled from a torus.

(b) A point cloud sampled from a sphere.

Fig. 2.9 Two point clouds sampled from (a) a torus and (b) a sphere. The point clouds were generated from Tadasets package [50]



(a) Persistent Diagram of a torus.

(b) Persistent Diagram of a sphere.

Fig. 2.10 Persistence diagrams for (a) a torus and (b) a sphere. Each point represents a topological feature, with its coordinates indicating the birth and death scales. Blue points (β_0) represent connected components, orange points (β_1) correspond to loops, and green points (β_2) represent cavities or voids.

In this study, persistent homology serve to compute descriptors of underlying homology patterns within gene expression datasets. By quantifying connected components (β_0), loops (β_1), and voids (β_2), they offer a robust framework for identifying topological features that may correlate with specific biological phenomena. These features are utilized to identify biomarkers and incorporated into downstream machine learning models, improving the accuracy of clinical outcome predictions.

2.11 Topological Descriptors

Persistence diagrams are essential tools in TDA, offering a way to represent the topological information. However, since they consist of variable-length collections of points, they are difficult to integrate directly into machine learning models, which typically require fixed-length vector inputs. To address this issue, researchers have developed representations called topological descriptors, which employ vectorization techniques to translate persistence diagrams into forms suitable for computation and analysis. Among the various methods available, persistence landscapes stand out for their robustness, and compatibility with statistical and machine learning frameworks [51, 52].

2.11.1 Persistent Landscapes

Persistence landscapes, introduced by Bubenik et al. [51], transform persistence diagrams into a collection of piecewise-linear functions. This transformation converts topological features into a vectorized format, making them compatible with statistical analysis and machine learning applications.

For each point (b,d) in a persistence diagram, a corresponding tent-shaped function (or "hat" function) is defined as:

$$\lambda_i(t) = \begin{cases} t-b, & \text{if } b \leq t \leq \frac{b+d}{2}, \\ d-t, & \text{if } \frac{b+d}{2} \leq t \leq d, \\ 0, & \text{otherwise.} \end{cases}$$

Where:

- b: Birth time of the topological feature
- d: Death time of the topological feature
- t: Parameter across the real line

Each function peaks at the midpoint between birth and death times and tends linearly to zero at both points. The collection of these hat functions forms a persistence landscape, denoted as:

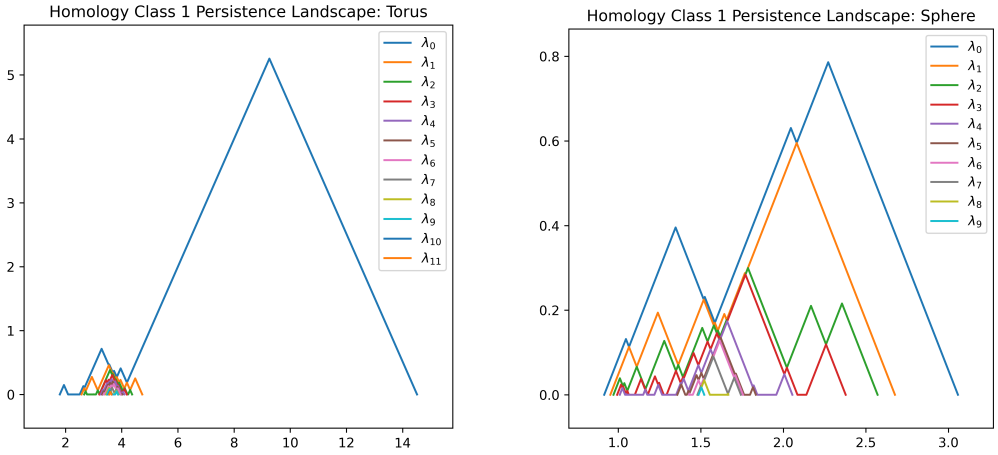
$$\Lambda = (\lambda_1(t), \lambda_2(t), \ldots)$$

These functions can be discretized into finite-dimensional vectors suitable for input into standard machine learning algorithms. Figure 2.11 illustrates persistence landscapes for β_1(loop-based) features of a torus and a sphere. Notice the substantial differences:

1. **Torus** (Figure 2.11a): A single dominant layer λ_0 extends over a large scale, reflecting a robust loop (β_1) that persists through much of the filtration. Higher layers ($\lambda_1, \lambda_2, \lambda_3 \ldots$) are less pronounced, indicating shorter-lived loops or noise.

2. **Sphere** (Figure 2.11b): In contrast, the sphere's β_1 features appear as small, overlapping peaks across all layers, with heights below 0.8 for all λ's. These short-lived loops are typically attributed to noise rather than a persistent topological structure.

Overall, the torus exhibits a robust and persistent β_1 feature, while the sphere's landscape is dominated by transient, short-lived features, reflecting the distinct topological properties for β_1 of the two shapes.



(a) Persistent Landscape of a torus.

(b) Persistent Landscape of a sphere.

Fig. 2.11 Persistence landscapes for β_1 of (a) a torus and (b) a sphere. Note the scale of the x- and y-axes.

In this thesis, we utilize persistence landscapes to generate a topological fingerprint for each patient. These topological fingerprints will be used as input features in downstream machine learning models to predict for cancer type, cancer stage and treatment response.

Chapter 3

Literature Review

Building upon the theoretical framework established in the preceding chapter, this literature review explores cancer biology and RNA-Seq data, with an emphasis on biomarker discovery and clinical outcome prediction. By situating this study in the broader scientific context, we aim to demonstrate TDA's potential to advance biomedical data analysis and contribute to improved biomarkers and clinical predictions.

3.1 Biology of Cancer

According to the World Health Organization (WHO), cancer is the leading cause of death globally, accounting for approximately 10 million deaths in 2020 alone [53]. Despite decades of intensive research, cancer remains a complex and multifaceted disease with significant implications for global health. Therefore, understanding the biological basis of cancer is essential for developing effective diagnostic tools and therapeutic interventions.

Cancer is caused by changes to genes that control how our cells function, particularly how they grow and divide. These genetic alterations can be inherited through germline mutations or acquired over a person's lifetime due to errors in deoxyribonucleic acid (DNA) replication or exposure to carcinogens such as tobacco smoke, ultraviolet radiation, or certain chemicals. These mutations disrupt the normal regulatory mechanisms of the cell cycle, leading to uncontrolled proliferation [54].

At the molecular level, the central dogma of molecular biology outlines the flow of genetic information from DNA to ribonucleic acid (RNA) to functional proteins [55]. This process is tightly regulated in healthy cells to ensure proper cell function, growth, and programmed cell death (apoptosis). Mutations in key genes result in gain/loss of function to oncogenes, tumor suppressor genes, and DNA repair genes—can lead to losing control over these processes. For example, the activation of oncogenes can promote excessive cell division, while the inactivation of tumor suppressor genes removes critical checks on cell growth [56]. Additionally, epigenetic modifications, such as

DNA methylation and histone modifications, can alter gene expression without changing the DNA sequence, further contributing to cancer progression [56].

Furthermore, cancer development is also heavily influenced by the tumor microenvironment. This microenvironment includes surrounding blood vessels, immune cells, fibroblasts, signaling molecules, extracellular matrix and the cancer stem cell populations. Interactions between cancer cells and the tumor microenvironment drive processes such as angiogenesis, invasion, and metastasis, which influencing therapy response [57].

With that being said, early detection and characterization of cancer is important, as it can significantly improve patient outcomes by allowing for timely intervention. Cancer detection begins with screening tests or assessments prompted by symptoms. Screening methods may include imaging techniques such as mammography for breast cancer, computed tomography (CT) scans, and colonoscopy for colorectal cancer [58]. Blood tests and liquid biopsy can also play a role in detection, measuring levels of specific biomarkers that may indicate the presence of cancer. Once cancer is detected, the next phase is obtaining a solid tumor biopsy to categorize the cancer. This process includes determining the malignancy, histological subtype, tumor size, and stage. Characterization is achieved through a combination of advanced imaging, morphological assessments via histology or cytology, and biomarker identification [58]. All these approaches aim to provide a comprehensive understanding of the cancer's nature, guiding personalized treatment strategies and improving prognosis.

Whilst traditional cancer screening and diagnosis has curbed some of the burden of cancer, advancements in genomic sequencing and omics technology (e.g., transcriptomics, proteomics) has the potential to revolutionize the field. These methods generate vast datasets, enabling the discovery of novel biomarkers and therapeutic targets. However, the complexity and scale of these datasets necessitate advanced analytical techniques, such as machine learning algorithms and TDA, to extract meaningful insights [6, 26].

3.1.1 Hallmarks of Cancer

Building on the foundational principles of cancer biology, Hanahan and Weinberg proposed the Hallmarks of Cancer, outlining the fundamental traits acquired during cancer's multi-step development [59, 60]. These hallmarks provide a unifying framework for understanding tumorigenesis and have been instrumental in guiding cancer research. The original six hallmarks include sustaining proliferative signaling, evading growth suppressors, resisting cell death (apoptosis), enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis [59–61]. (Figure 3.1). Later, two emerging hallmarks were added namely reprogramming energy metabolism and evading immune destruction [59, 61].

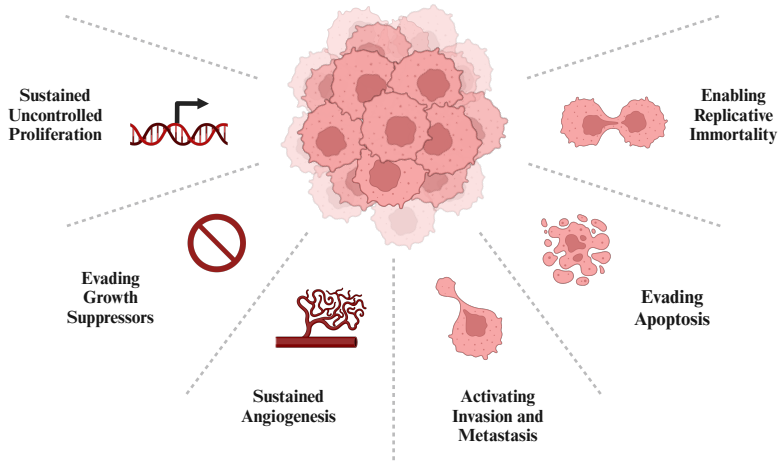


Fig. 3.1 Key hallmarks of cancer include sustained proliferative signalling, evasion of growth suppressors, angiogenesis, invasion and metastasis, evasion of apoptosis, and replicative immortality. Figure was created in BioRender.com

3.1.1.1 Sustaining Proliferative Signaling

Proliferation refers to a rapid reproduction of a cell, part or organism. This leads to tissue growth or an expansion of cell populations. This process is tightly regulated in health tissues by signals that promote or inhibit cell division as needed. However, in cancer, cells acquire the ability to sustain uncontrolled proliferation, meaning they continuously multiply without responding to normal growth controls. This occurs through mutations in oncogenes and the inactivation of tumor suppressor genes, which allow cancer cells to bypass regulatory checkpoints and evade signals that would normally stop their growth [60].

3.1.1.2 Enabling Replicative Immortality

Cancer cells can have limitless replication potential, which is known as replicative immortality. This contrasts the behavior of normal cells which undergo a limited number of growth and division cycles before encountering senescence, a non-dividing state, or crisis, which leads to cell death. However, cancer cells circumvent these limitations by maintaining the length of their telomeres, the protective caps at the ends of chromosomes that shorten with each division. This is primarily achieved via the hyperexpression of the telomerase enzyme, which extends telomeres and preserves chromosomal stability. By overriding this natural mechanism of cell number regulation, cancer cells evade processes such as necrosis and apoptosis, further driving uncontrolled proliferation [60].

3.1.1.3 Evading Growth Suppressors

Growth suppressors, also known as tumor suppressor genes, regulate cell proliferation by acting as "brakes" that inhibit unchecked growth. Well-documented tumor suppressors commonly mutated in cancer include Retinoblastoma (RB) and TP53, enabling cancer cells to evade growth suppression. The RB and TP53 proteins play key roles in regulating cellular proliferation, senescence[1], and apoptosis. RB integrates signals to control cell cycle progression, while TP53 responds to intracellular stress by halting growth or triggering cell death when conditions are abnormal. Despite their critical function, cancer cells often bypass these suppressors through genetic mutations or pathway inactivation, leading to unchecked proliferation [60].

3.1.1.4 Activating Invasion and Metastasis

Invasion and metastasis are processes by which cancer cells spread from their original site to other parts of the body. Invasion involves cancer cells penetrating and spreading into surrounding tissues, while metastasis is the disseminating of cancer cells through the bloodstream or lymphatic system to distant organs, where they establish secondary tumors. Both processes are driven largely by epithelial-mesenchymal transition (EMT). EMT induces morphological changes in cancer cells, disrupting cell-cell interactions and enabling mobility [60]. Notably, many sources indicate metastasis accounts for approximately 66-90% of cancer-related deaths, making it the leading cause of cancer mortality [62, 63].

3.1.1.5 Evading Apoptosis

Apoptosis is the process of programmed cell death, a natural and controlled mechanism that allows cells to die in a way that benefits the organism. Evading apoptosis refers to the ability of cancer cells to avoid programmed cell death, allowing them to survive and continue to proliferate despite signals that generally trigger cell death [60].

3.1.1.6 Sustained Angiogenesis

The continuous delivery of oxygen and nutrients, along with the elimination of metabolic waste, is critical for the survival and optimal functioning of cells. Angiogenesis refers to the physiological process of forming new blood vessels from pre-existing vessels, which plays an important role in tissue growth, repair and survival. In the context of cancer, angiogenesis plays a pivotal role by supplying the tumor with the necessary oxygen and nutrients to sustain its growth and progression. Tumors stimulate this process by secreting pro-angiogenic factors, such as vascular endothelial growth factor (VEGF), which promote the growth of new blood vessels toward the tumor mass [60].

[1] Senescence is a permanent state in which cells stop dividing in response to stress or damage, serving as a protective mechanism to prevent uncontrolled growth.

3.2 Biological data

3.2.1 RNA Sequencing

With the biology and the hallmarks of cancer defined, developing analytical methods to explore the genetic landscape is essential for addressing the complexities of cancer. Genes, the fundamental units of inheritance, contain the information that is necessary to determine an organism's physical and biological traits. Passed from parent to offspring, genes encode information for producing molecules, primarily proteins, that perform vital cellular functions [64]. This information is carried within a gene's nucleotide base pairs, which is composed of nucleotides namely adenine (A), thymine (T), cytosine (C), and guanine (G). This information can be decoded using high-throughput sequencing technologies, such as DNA microarrays and next generation sequencing technologies such as RNA sequencing (RNA-Seq). These technologies enable the comprehensive profiling of gene expression data, allowing researchers to measure thousands of genes simultaneously and uncover molecular patterns associated with cancer [56, 65].

Older technologies such as DNA microarrays, consist of thousands of DNA fragments (spots) affixed to a glass or silicon chip, with each spot representing a different gene. This setup enables the simultaneous analysis of numerous genes in a single experiment. However, microarrays can only detect transcripts that bind to pre-existing spots, making them dependent on known genomic and transcriptomic data. Additionally, they typically detect only two-fold expression changes, which may cause them to miss subtler yet biologically significant variations [66].

Newer technologies, such as RNA-Seq, analyze the complete set of RNA transcripts within a cell or tissue. By converting RNA into complementary DNA (cDNA) and sequencing it, researchers can quantify gene expression and uncover novel transcripts, providing a snapshot of cellular activity and gene regulation. The RNA-Seq protocol involves fragmenting and tagging the cDNA with adapters, followed by amplification and sequencing, where each DNA base emits a unique signal during synthesis. Compared to microarrays, RNA-Seq is more sensitive and captures both known and novel transcripts, including splice variants, enabling the detection of subtle expression changes at lower thresholds [66].

While short-read RNA-Seq (e.g., Illumina-based) remains the most widely used approach due to its mature protocols, high throughput, and cost-effectiveness, it has limitations in resolving long-range transcript structures and complex splicing events. In contrast, long-read RNA-Seq (e.g., Oxford Nanopore, PacBio) is now considered the next generation of RNA sequencing, offering full-length transcript information, enhanced isoform detection and reduced sequence bias [67, 68]

3.2.1.1 Challenges in RNA-Seq

RNA-Seq data analysis brings formidable challenges, particularly in biomarker discovery and classification. One major issue is the vast number of molecular features, such as the thousands of genes that could serve as potential biomarkers or important features for classification. Traditional molecular and cell biology research focuses on elucidating isolated and specific biological pathways/processes rather than considering all genes. This limits the overall understanding of the entire biological system. As such, it is of growing importance to consider all genes when identifying biomarkers, as often time, the combined expression patterns of multiple genes can offer better prognostic power than individual genes. However, as the feature set grows, identifying the optimal subset for reliable prognosis becomes complex, a challenge often termed as the curse of dimensionality. For instance, consider a panel of 10 genes, which can result in over 10! = 3.6 million combinations. This complexity also hampers data visualization and interpretation, obscuring critical patterns in gene expression profiles [69].

Moreover, RNA-Seq data often presents additional challenges, such as sparse data matrices, technical noise, and batch effects, as well as both biological and technical variability [70, 71]. Addressing these challenges is crucial for extracting meaningful insights from RNA-Seq data.

3.3 Methodologies

Analyzing RNA-Seq data, particularly in oncology, demands robust methodologies that extract meaningful insights from high-dimensional, sparse, noisy and heterogeneous datasets. Statistics, machine learning, and bioinformatics provide essential tools and frameworks for this task, enabling researchers to detect patterns, build predictive models, and interpret results. However, these methods face overlapping limitations when dealing with the inherent complexity of RNA-Seq data (See Section 3.2.1.1). In this section, we discuss the key limitations of these methodologies and illustrate how TDA emerges as a complementary framework for addressing these challenges.

3.3.1 Challenges of Statistics

Statistics has been instrumental in advancing biological research, by providing us the tools for hypothesis testing, pattern identification, and data summarization. It underpins many bioinformatics workflows and machine learning algorithms. However, statistical methods encounter significant limitations when applied to RNA-Seq datasets.

1. **Deceptive Summary Statistics**

   Statistical methods often rely on summary metrics such as mean μ standard deviation σ, and Pearson's correlation r to describe data. However, these metrics may fail to capture the underlying geometric and structural relationships within complex datasets. This limitation is illustrated by the Datasaurus Dozen, introduced by Matejka et al. [72], where datasets with

identical statistical properties (μ, σ, r) exhibit vastly different visual patterns and distributions. This phenomenon is depicted in Figure 3.2. This limitation is further compounded by the fact that these summary statistics form the backbone of many statistical and machine learning models. These statistics are fundamental to Gaussian models, linear, multivariate, and logistic regression, principal component analysis (PCA), k-nearest neighbors (k-NN), and k-means clustering. Additionally, they are commonly employed in data normalization techniques for machine/deep learning models like decision trees and neural networks, highlighting their pervasive role in statistics and machine learning. The phenomenon illustrates the idea that summary statistics alone cannot discern these differences, as it overlooks the geometric and topological properties that are critical to understanding the true nature of the data.

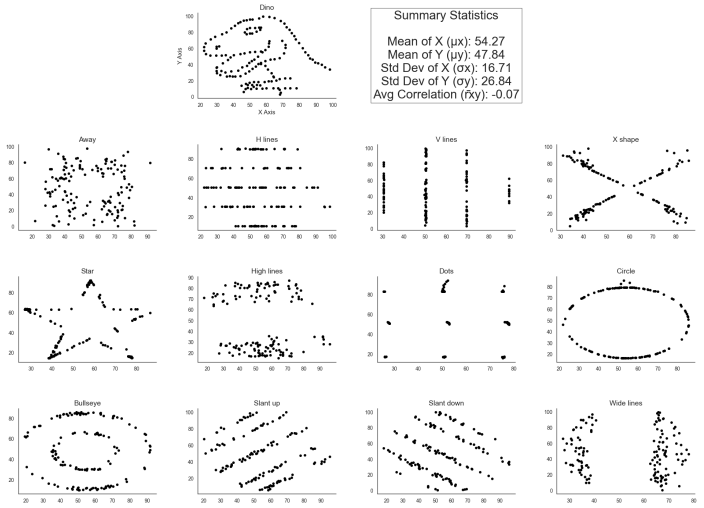


Fig. 3.2 The Datasaurus Dozen showcasing a collection of 13 datasets that share nearly identical summary statistics—mean (μ), standard deviation (σ), and Pearson correlation (r) yet vary dramatically in their distributions and shapes. This highlights a key limitation of relying solely on summary statistics underscoring the importance of examining the shape of data. The data points were retrieved from [73]

2. **Correlation or Causation**

   Another challenge is that statistical methods are prone to detecting correlation without establishing causation. Correlation alone does not imply causation, and spurious correlations can arise, creating misleading conclusions about the strength or nature of observed connections [74]. Additionally, correlation-based approaches may fail to detect non-linear relationships, limiting their capacity to capture the intricate, multidimensional structures inherent in complex biological datasets [75].

3. **Outliers**

   Lastly, outliers and extreme values, arising from biological variability or technical noise, can distort statistical analyses and lead to unreliable conclusions [76]. These anomalies are

especially problematic in datasets like gene expression profiles, where subtle biological signals are critical.

3.3.2 Challenges of Machine Learning

Machine learning, a branch of AI, focuses on the developing algorithms capable of learning from data and generalizing to new, unseen cases, thus performing tasks without explicit programming. These algorithms generalize insights from data to unseen cases, enabling tasks such as disease classification, biomarker discovery, and clinical outcome prediction. However, like statistics there are significant challenges regarding machine learning in complex biological datasets.

1. **Data Representation and Feature Selection**

   In machine learning, data representation and feature selection play a crucial role in optimizing model performance and interpretability [77, 78]. Data representation involves organizing complex information to make patterns and relationships meaningful to algorithms [77]. Meanwhile, feature selection is the strategic process of isolating the most relevant aspects of the data and filtering out noise and redundancy to optimize model accuracy and efficiency [78]. This alludes to a familiar maxim in computer science, "garbage in, garbage out", emphasizing that the model's performance hinges on data quality. Flawed input inevitably leads to flawed results [79]. This becomes a significant challenge for high-dimensional sparse data such as RNA-Seq data. Careful representation and selection of molecular features are essential, as each chosen feature contributes to the model's ability to learn and generalize effectively.

2. **Black Box Problem**

   Advanced machine learning models, particularly deep neural networks, suffer from poor interpretability. As these models grow in complexity, with layers upon layers of interconnected parameters, it becomes virtually impossible to trace the exact steps through which decisions are made. This phenomenon, known as the *black box problem*, limits our understanding of the logic used by these algorithms to generate results. Without transparency, the logic followed by neural networks remains largely unexplainable, which can undermine trust in their applications, especially in fields where interpretability is essential. Studies indicate that limited transparency and understanding of the inner workings of the machine learning models are key factors impeding their integration into clinical settings [80]. For instance, convolutional neural networks (CNNs) are state-of-the-art in image recognition, however at the same time are notoriously difficult to interpret, thus contributing to the limited adoption in a clinician setting. Without explainability, clinicians and stakeholders may hesitate to adopt models that cannot provide clear insights into their decision-making processes.

3. **Reliance on Summary Statistics**

Machine learning models often rely on statistical descriptors such as mean (μ), standard deviation (σ), and correlation (r) for normalization and initial analysis as discussed using the Datasaurus Dozen example in Section 3.3.1. However, by relying heavily on these metrics, machine learning models inherit the limitations of summary statistics, which can be misleading and often overlook geometric and topological properties. Thus, reducing the model's capacity to represent and interpret data in a way that aligns with its true underlying complexity [73].

4. **Small Sample Sizes**

   Another significant challenge is imposed by small sample sizes. Many advanced machine learning methods are inherently data-hungry; their performance improves significantly with larger datasets. When data is scarce, which is often the case in fields such as genomics or rare disease studies, these models struggle to generalize effectively, leading to diminished accuracy and robustness [81]. With small sample sizes, models are prone to overfitting, where the model memorizes the training data instead of generalizing patterns. This results in poor performance on unseen data [82, 83]. Even the most sophisticated algorithms cannot reach their full predictive potential without a sufficient number of samples, emphasizing the need for strategies to maximize insights from limited data.

These limitations underscore the need for innovative methods, like TDA, which can address structural complexity and enhance machine learning models by revealing hidden geometric and topological patterns in the data.

3.3.3 Bioinformatics

Bioinformatics serves as the computational backbone for modern biological research, offering essential tools and methodologies to store, analyze, and interpret vast and complex datasets. It has become indispensable in tasks such as genome assembly, annotation, disease gene identification, and personalized medicine. However, as an interdisciplinary field that integrates statistical techniques and machine learning algorithms, bioinformatics inherits key limitations from both disciplines mentioned above (See Section 3.3.1 and 3.3.2).

These limitations across statistics, machine learning, and bioinformatics highlight the pressing need for innovative methodologies capable of addressing structural complexity, interpretability, and noise in RNA-Seq datasets. TDA, more specifically persistent homology emerges as a powerful complementary framework, offering a multi-scale perspective that captures the intrinsic shape and topology of data, enabling deep insights into complex biological patterns.

3.3.4 Topological Data Analysis (TDA)

TDA has shown promising results in analyzing high-dimensional datasets across a wide range of domains, including Natural Language Processing (NLP) [84, 85], neuroscience [86, 87], time-series

modeling [88, 89], financial modeling [90, 91]. More importantly, it has recently emerged as a powerful and complementary approach for analyzing complex biological datasets in oncology [14, 22, 23, 26, 92–97].

While statistical, machine learning, and bioinformatics methods have significantly advanced our ability to analyze biological data, these approaches alone often fall short in addressing the inherent complexity of biological datasets, as discussed above (See Section 3.3). Challenges faced by statistical methods include deceptive summary statistics, spurious correlations, and sensitivity to outliers. Moreover, machine learning methods grapple with issues such as data representation, feature selection, interpretability, reliance on summary statistics, and limitations posed by small sample sizes.

Persistent homology overcomes the pitfalls of deceptive summary statistics and spurious correlations by shifting the analytical focus to the geometric and topological properties of the data. By analyzing the persistence of topological features across multiple scales, it identifies patterns that remain stable over a wide range of resolutions. This allows it to effectively differentiate meaningful structures from those that result from noise or outliers [34].

Furthermore, persistent homology evaluates features based on their persistence, where the longevity of topological structures across scales indicates their significance. This persistence-based criterion effectively captures stable topological features, highlighting the most relevant and informative patterns, even in scenarios with small sample sizes or limited data points [98].

Persistent homology also addresses the interpretability problem commonly encountered in machine learning models. By offering traceability back to the original data points that contribute to the formation of β numbers or homology groups, persistent homology ensures that each identified topological feature can be directly mapped to its corresponding data points. These topological descriptors can manifest as connected vertices (β_0), cycles (β_1), or voids (β_2) [31]. This traceability enhances transparency and interpretability [14].

As such, TDA serves as a powerful addition to the analytical toolkit, by dealing with some of the inherit challenges faced by statistics and machine learning. In this thesis, persistent homology will be applied to identify biomarkers using WGTDA and to predict clinical outcomes using RNA-Seq data.

3.4 Biomarker Discovery

A biomarker, or biological marker is defined as a characteristic that is evaluated objectively to indicate normal biological processes, pathogenic processes, or pharmacologic responses to therapeutic interventions [69]. Biomarkers can be found in various biological materials such as DNA, RNA, blood, urine, and tissue, and they provide valuable insights into an individual's health [99]. In

oncology, a biomarker may refer to any measurable indicator of the risk of cancer, cancer occurrence or patient outcome [100]. Biomarkers are crucial for translating molecular understanding into clinical applications, ultimately enabling earlier and more precise disease diagnoses, deeper understanding of disease mechanisms, and improvements in patient care [14, 99].

Among the various types of biomarkers, a prognostic biomarker offers information about the likely progression or outcome of a disease in an untreated individual [99, 101]. Furthermore, prognostic biomarkers can be categorized into two distinct groups: those that provide insights into the likelihood of recurrence in patients undergoing curative treatment and those that correlate with the duration of progression-free survival in patients with metastatic disease [99].

This thesis employs WGTDA to identify biomarkers from RNA-Seq data for three primary tasks: cancer type, cancer stage, and treatment response. Furthermore, this work aims to introduces the development of a custom visualization tool designed to facilitate the analysis and interpretation of topological features. These tasks directly address research aims 1.2.3.1 and 1.2.3.2 which state:

- *Identify prognostic biomarkers in gene expression data using WGTDA for cancer type, cancer stage and treatment response.*
- *Develop a visualization framework to represent topological features in gene expression as a complex network*

We aim to identify these biomarkers by leveraging WGTDA's topological perspective in uncovering structural patterns in co-expressed gene networks through the use of *betti* numbers and homology groups [14, 15]. By leveraging the visualization tool and the persistent *betti* features identified through WGTDA, we will conduct external validation using independent datasets and literature. This validation aims to rigorously evaluate whether WGTDA demonstrates clinical relevance and robustness as a bioinformatics tool for biomarker discovery.

3.5 Clinical Outcome Prediction

Predicting clinical outcomes refers to the process of estimating the probable course, progression, and eventual response of a patient's disease. Accurate outcome prediction is important for guiding treatment plans, assessing risk and improving survival rates. The goal of outcome prediction is to provide clinicians with a comprehensive assessment that informs medical decisions, aiming to optimize patient care and minimize potential adverse effects.

Several approaches have been developed to predict clinical outcomes, each contributing to our understanding and management of disease progression:

1. **Clinical analysis**

   Traditional methods relied heavily on clinical analyses, where observable parameters such as tumor size, location, histological grade, and demographic factors like age, gender, and lifestyle—served as primary indicators [102]. This approach is extremely valuable but often suffers from limited insights, as it does not account for the underlying molecular characteristics that drive disease behavior.

2. **Molecular Biomarkers**

   In recent years, molecular biomarkers have become instrumental in clinical outcome prediction. These include genes, proteins, and specific genetic mutations, which reveal critical details about a tumor's behavior and its likely response to treatment. Techniques such as immunohistochemistry (IHC) and polymerase chain reaction (PCR) allow for the detection and analysis of these biomarkers. PCR gained widespread recognition during the COVID-19 pandemic for its role in real-time viral RNA detection, which exemplifies how molecular analysis can enhance diagnostic and predictive accuracy [103].

3. **Computer-Aided Diagnosis**

   Computer-Aided Diagnosis (CAD) systems assist clinicians in analyzing both imaging and non-imaging data, particularly in radiology, where they aid in tumor detection and staging using mammograms, CT (computed tomography) scans, and MRI (magnetic resonance imaging) [104]. Acting as a "second opinion" for radiologists, CAD highlights abnormal structures and subtle patterns, enhancing diagnostic and disease progression predictions.

4. **Electronic Health Records (EHR)**

   The analysis of EHR offers valuable historical data, providing insights into patient outcomes based on profiles with similar characteristics [105]. With detailed records of medical history, treatment regimens, and clinical outcomes, EHR enables comparative analysis, enhancing the accuracy of outcome predictions and supporting evidence-based medical decisions.

5. **Machine Learning and Deep Learning**

   With the rise of large-scale, complex biological data, machine learning and deep learning have become essential for clinical outcome prediction. These models uncover hidden patterns in both structured and unstructured data, including RNA-Seq, single-cell RNA-Seq, and medical imaging data, enabling accurate, data-driven decisions in healthcare [106].

In this thesis, we leverage TDA descriptors with machine learning models to enhance predictive accuracy. By generating unique topological fingerprints of different co-expression measures for each patient, we aim to classify RNA-Seq data for cancer type, cancer stage, and treatment response. This task directly addresses the research aims 1.2.4.1 and 1.2.4.2:

- *Evaluate the impact of integrating topological descriptors on the predictive performance of machine learning models for clinical outcomes.*
- *Compare different co-expression measures in constructing the simplicial complex and assess their influence on model performance.*

By evaluating these aims across tasks of increasing complexity, we aim to demonstrate the superiority of TDA descriptors over traditional data representations and identify the most effective co-expression measures for topological descriptors in downstream machine learning tasks.

3.6 Thesis Tasks

This thesis is structured around three tasks leveraging TDA and machine learning to enhance clinical outcome classification and biomarker discovery, each task increasing in complexity and difficulty. Task 1 (cancer type) focuses on distinguishing different cancer types. Task 2 (cancer staging) classifies patients into distinct stages. Lastly, task 3 (treatment response) aims to identify biomarkers and classification linked to treatment resistance and sensitivity.

3.6.1 Task 1: Cancer Type

Cancer is a heterogeneous disease categorized by its tissue or organ of origin, which impacts behavior, treatment response, and prognosis. In this task, we leverage data from The Cancer Genome Atlas (TCGA) project, a comprehensive resource that provides molecular and clinical insights across various cancer types. Specifically, we focus on biomarker discovery and classification for three cancer types: sarcomas (TCGA-SARC), esophageal carcinomas (TCGA-ESCA), and paragangliomas/pheochromocytomas (TCGA-PCPG).

- **Sarcomas (SARC)**

  Sarcomas are aggressive cancers originating in connective tissues such as bone, muscle, or fat. They are highly heterogeneous, posing a significant challenge in treatment due to their tendency to reoccur and metastasize. These cancers are often associated with poor outcomes, requiring precise diagnostic and therapeutic strategies [107, 108].

- **Esophageal carcinomas (ESCA)**

  Esophageal carcinomas is a growth of cells originating in the esophagus. They are classified into two main histologic subtypes: squamous cell carcinoma and adenocarcinoma. This cancer type is often diagnosed at an advanced stage, as symptoms typically emerge late in its progression. Consequently, esophageal carcinoma has a low five-year survival rate of 29.3% for cases diagnosed between 2005 and 2011, highlighting the urgent need for early detection and effective intervention [109, 110].

- **Paragangliomas and Pheochromocytomas (PCPG)**

  Paragangliomas and Pheochromocytomas are rare cancer located in the adrenal glands. The adrenal glands are two small organs that are situated on top of each kidney, and they help with bodily functions including but not limited to blood pressure and metabolism. When they develop in the center of the adrenal gland, they are referred to as pheochromocytomas [111]. These cancers are extremely rare, with an incidence of 2–8 cases per million annually worldwide. Approximately 20% of all cases occur in children [112].

3.6.2 Task 2: Cancer Staging

Cancer staging refers to the extent of the cancer, which is determined by factors such as tumor size, the time since symptoms first appeared, and the spread of the cancer. Survival likelihood is closely linked to the cancer's stage at diagnosis [113, 114]. The widely used TNM staging system classifies cancers based on three main criteria: the size and extent of the primary tumor (T), the involvement of nearby lymph nodes (N), and the presence of distant metastases (M) [113, 115, 116]. Together, these factors provide a comprehensive assessment of the cancer's progression, guiding treatment decisions and informing prognosis.

Within the TNM system, cancers are grouped into stages that reflect their progression and spread, as described by the TNM Classification from The National Library of Medicine [116, 115, 114]:

- **Stage 0:** Also known as carcinoma *in situ*, this earliest stage of cancer consists of abnormal cells that are confined to their original location. Although non-invasive, these cells have the potential to become cancerous and may eventually spread into nearby tissue [114].
- **Stage I:** Often called to as early-stage cancer, this stage is characterized by small, localized tumors with minimal risk of spreading to nearby tissues or lymph nodes. Treatment, such as surgery or radiation, is typically very effective, and the prognosis is generally positive [114].
- **Stage II:** In this stage, the cancer is larger or has begun to invade nearby tissues but still has not spread to other parts of the body. Stage II requires more intensive treatment due to its greater size or involvement of nearby tissue, but it has not spread beyond the primary site [114].
- **Stage III:** Referred to as locally advanced cancer. At this stage the cancer has either increased in size, spread to nearby lymph nodes, or affected surrounding tissue. This stage often requires a combination of treatments, such as surgery, radiation, and chemotherapy, as the disease has become more complex to manage [114].
- **Stage IV:** The most advanced stage which is also known as metastatic cancer. In this stage the cancer has spread (metastasized) to distant parts of the body. At this stage, the focus shifts toward controlling the spread, managing symptoms, and enhancing quality of life, as curative treatments are often challenging [114].

Although the TNM system provides a standardized method for classifying cancer stages, identifying biomarkers and accurately distinguishing between stages remains challenging. The AJCC Cancer Staging Manual, first published in 1977, introduced a standardized approach to cancer staging while acknowledging its limitations as "not an exact science". The editors recognized that cancer staging would need to evolve as new information on etiology, diagnosis, and treatment became available [117]. A key challenge has been differentiating intermediate stages, particularly stages II and III, which often overlap in characteristics. In response to such challenges, the Eighth Edition of the AJCC Manual builds on anatomic staging by incorporating biological and molecular markers, reflecting a shift towards on precision medicine, where staging is tailored to individual patient profiles rather than relying solely on population-based approaches [117].

Building on this progression toward individualized care, this thesis, identify biomarkers and classifies patients according to their cancer stage. The biomarkers and topological fingerprints aim to capture patient-level variations crucial for precise stage differentiation, even between challenging stages like II and III. Using data from TCGA-HNSC (Head and Neck Squamous Cell Carcinoma), which spans cancer stages I through IV, we aim to classify patients accurately across these distinct stages.

3.6.3 Task 3: Treatment Response

Predicting treatment response is crucial for personalizing cancer therapies and improving patient outcomes. Arguably, treatment response is the most challenging task as compared to cancer type and staging as it is complicated by intrinsic factors, such as genetic mutations, and extrinsic factors, including the tumor microenvironment. Chemotherapy, a cornerstone of cancer treatment, uses drugs like cisplatin, paclitaxel, doxorubicin, and 5-fluorouracil (5-FU) to target rapidly dividing cells. Similarly, radiation therapy, often combined with agents like cisplatin or cetuximab, induce DNA damage in cancer cells to maximize therapeutic effects. However, resistance mechanisms often limit the efficacy of these treatments, making patient-specific response prediction particularly complex.

Key mechanisms contributing to treatment resistance include:

- **Epithelial-Mesenchymal Transition (EMT):** where epithelial cells lose adhesion and polarity, which causes cancer cells to acquire migratory and invasive properties, leading to increased resistance to chemotherapy and immunotherapy [59, 60, 118, 119]
- **Enhanced vascular endothelial growth factor (VEGF):** Elevated VEGF promotes angiogenesis, supporting tumor growth and reducing the efficacy of VEGF-targeted therapies [120, 121].
- **Hypoxia:** is a reduction in tissue oxygen levels, which promotes aggressive tumors, reduces therapies effectiveness, and worsens clinical outcomes [122–124]

- **Metabolic Reprogramming:** is where alterations in glucose and lipid metabolism provide cancer cells with the energy to survive and proliferate under treatment stress [125, 126].

In this thesis, we leverage TDA to identify biomarkers and improve treatment response prediction by distinguishing resistant/non-responders from sensitive/responders. To achieve this, we utilize data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC), a large-scale initiative that provides high-quality proteomic and genomic data to advance cancer research. Ultimately, we aim to uncover robust biomarkers and improve clinical outcome prediction on how patients will respond to therapy.

3.7 Related Works

This section reviews the body of research that has shaped our understanding of TDA in gene expression data, emphasizing its role in biomarker discovery and clinical prediction. We structured this review into two parts namely, biomarker discovery and clinical outcomes to examine key studies in these domains and situate this thesis within the broader field of computational biology and TDA. Our goal is to demonstrate how persistent homology has advanced our ability to interpret RNA-Seq data, uncovering novel biological insights while addressing recurring challenges that this thesis seeks to overcome.

3.7.1 Biomarker Discovery

One of the earliest applications of TDA in biology as a proposed biomarker tool was the use of the Mapper algorithm introduced by Singh et al. [127]. The Mapper algorithm is a TDA method designed to simplify and visualize high-dimensional data by creating a graph-based representation. It does this by mapping data into a combinatorial structure, known as a simplicial complex, which captures the shape and connectivity of the data, thus revealing clusters, loops, and other significant structures in a user-defined resolution. The Mapper algorithm was employed to identify a distinct subgroup of Estrogen Receptor-positive (ER+) breast cancers. This subgroup is characterized by high expression levels of the c-MYB gene and low expression of innate inflammatory genes. Remarkably, patients within this group demonstrated 100% survival rates with no observed metastasis [128]. Although this thesis centers on persistent homology, this early application demonstrates the TDA usefulness and versatility in oncology.

Building upon the successes of Mapper, persistent homology has emerged as a useful tool for biomarker discovery. In contrast to Mapper which constructs a graph-based representation of the data, persistent homology provides a richer multi-scale view by characterizing topological spaces and structures across multiple dimensions, capturing topological features as they form and die over varying thresholds.

This approach is demonstrated in a study by Masoomy et al. [94], which used persistent homology to analyze β_1 and β_2 topological interactions in weighted networks of normal and cancerous gene expression data. Their findings reveal distinctive structural patterns within cancerous gene regulatory networks (GRN). More specifically an increase in persistent one-dimensional loops (β_1) and a decrease in higher-dimensional voids (β_2), compared to healthy GRNs. By examining the *betti* curves at β_1 and β_2 for cancerous patients suggests the cancer cells rely on certain pathways at the network level. This is a biological phenomenon called oncogene addiction. These insights showcase persistent homology's potential as a powerful tool in distinguishing cancerous from healthy cellular networks, providing a new avenue for biomarker discovery.

Another prominent application of persistent homology for biomarker discovery is presented in the study by Abdullahi et al. [95], where persistent homology was employed to identify critical pathways for the early detection of hepatocellular carcinoma (HCC). Topological interactions within RNA-Seq data were identified from peripheral blood of HCC patients and normal controls. The study found topological interactions consisting of genes implicated in key pathways for HCC pathology, such as the apelin, IL-17, and p53 signaling pathways. Moreover, a comparative analysis was conducted evaluating the enriched pathways identified by both topological interactions and differential expression analysis. Notably, while the IL-17 signaling pathway was identified by both methods, the HCC-related apelin signaling and p53 signaling pathways emerged exclusively through the persistent homology approach. These findings further contribute to the growing body of research emphasizing the role of TDA in uncovering biomarkers crucial to oncology.

Finally, a central study and tool for this thesis is WGTDA, a novel biomarker discovery framework developed in collaboration with my esteemed colleagues at IBM Research [14, 15]. Designed to bridge biology and topology, WGTDA enables the mapping of intricate topological interactions to specific gene signatures, thus reducing the ambiguity of which genes correspond to which topological features. This enhancement improves the interpretability and applicability of TDA-derived biomarkers. In the study Nyase et al. [14], WGTDA was applied to The Cancer Genome Atlas (TCGA) datasets for Breast Cancer (BRCA), Lung Adenocarcinoma (LUAD) and Colorectal Adenocarcinoma (COAD) and is compared to Weighted Gene Co-Expression Network Analysis (WGCNA), a data mining technique to identify hub genes by constructing gene co-expression networks and detecting biologically relevant gene modules [129]. WGTDA identified gene signatures with strong associations to survival, revealing key genetic markers linked to time-to-death outcomes. Kaplan-Meier (KM) survival analysis and Random Forest survival analysis demonstrated that WGTDA-derived gene signatures were more significant predictors of survival compared to those identified by WGCNA. Additionally, WGTDA identified disease-specific pathways, emphasizing key biological pathways relevant to each cancer type investigated.

Despite these advancements, three significant limitations persist in the application of TDA for biomarker discovery:

1. **Interpretability and Visualization Challenges:** There is an inherent ambiguity in mapping the abstract topological features back to the basic biological elements, such as genes. Without effective visualization tools, interpreting and exploring these complex topological interactions becomes difficult, limiting our understanding of their biological significance. Developing specialized visualization tools could be a key step in making TDA more accessible and widely adopted in computational biology.

2. **Biomarker Identification for Complex Phenotypes:** Many existing studies focus on distinct disease comparisons without accounting for complex phenotypic variations such as different stages or treatment response [14, 94, 95]. This gap in representation raises concerns about the broader applicability of persistent homology, as its effectiveness in capturing biomarkers within heterogeneous and homogeneous disease environments remains largely unproven.

Addressing these limitations, this thesis introduces several key points of novelty:

1. **Development of a Visualization Tool:** We propose a novel framework to visualize topological interactions as a complex network for WGTDA, providing deeper insights into the biological significance of these features. This directly tackles the challenge of interpreting topological information and seeks to answers research aim 1.2.3.2: *Develop a visualization framework to represent topological features in gene expression as a complex network.*

2. **Biomarker Identification for Complex Phenotypic Variations:** Unlike prior studies that focus on distinct disease comparisons, we apply WGTDA to uncover biomarkers associated with complex phenotypic variations, including cancer staging and treatment response. This directly addresses the research aim 1.2.3.1 *Identify prognostic biomarkers in gene expression data using WGTDA for cancer type, cancer stage and treatment response.* This seeks to determine the applicability of TDA and persistent homology in precision oncology by capturing nuanced disease characteristics.

3. **Validation of WGTDA as a Effective Framework for Biomarker Discovery:** Building upon previous work in Nyase et al. [14, 15], we aim to rigorously validate WGTDA, demonstrating its reliability and effectiveness as a biomarker discovery tool in TDA. This strengthens the evidence for its broader adoption in cancer and omics research.

By tackling these challenges, this thesis contributes to overcoming significant limitations in TDA for biomarker discovery. Our work enhances the interpretability of topological features through a novel visualization framework, validates WGTDA as a robust clinically relevant tool, and demonstrates its effectiveness in identifying biomarkers linked to heterogeneous disease phenotypes in cancer research.

3.7.2 Clinical Outcome Prediction

Recent studies have demonstrated the potential of integrating topological descriptors with machine learning techniques to enhance predictive power in various biomedical applications. One such study, conducted by Mandal et al. [93], explored the use of TDA in predicting phenotypes from gene expression data, specifically in the context of Parkinson's disease. The authors found that traditional machine learning approaches alone were insufficient for disease phenotype prediction, prompting the integration of TDA to improve classification outcomes. By utilizing persistent homology to extract topological features from transcriptomic data, they were able to generate topological summaries for each individual sample. These topological summaries were then integrated into machine learning models, resulting in significantly improved performance compared to standard machine learning techniques such as SVM and random forests.

In addition to this, Dey et al. [96] utilized persistent homology in gene expression data for phenotype prediction applying their methods to three datasets including dengue fever, bone marrow failure, and Crohn's disease. They introduced a unique approach by generating representative persistent cycles rather than relying solely on traditional barcodes, allowing for the identification of topologically relevant cohorts and gene expressions. This method significantly improved both shallow learning and deep learning-based classification tasks. Both studies underpin the importance of topological features as critical indicators in high-dimensional genomic data.

Furthermore, Mashatola et al. [22] expanded upon these approaches by optimizing the construction of the simplicial complex through the use of topological overlapping measures (TOM), a method traditionally employed in WGCNA to capture more robust and meaningful co-expression networks [129]. While correlation measures have been a common practice for building simplicial complexes in TDA, by replacing it with TOM, the study yielded a significant improvement in constructing the VR complex, which enhanced the identification of persistent topological features in cancer datasets. When the topological features are applied to downstream deep learning tasks for phenotype prediction between BRCA, LUAD, COAD/READ and Prostate Adenocarcinoma (PRAD) cancer types, TOM-based VR complexes achieved nearly 90% classification accuracy, outperforming distance correlation by approximately 15-20%. This study highlights the importance of selecting appropriate co-expression measures when generating topological descriptors from omics data. Furthermore, this study underscores the value of topological features as dependable indicators for phenotype prediction, revealing informative signals embedded in high-dimensional genomic datasets.

While TDA has shown considerable promise for clinical prediction, two notable limitations remain in its application for clinical outcome prediction:

1. **Predicting Complex Phenotypes Variations:** Similar to its challenges in biomarker discovery, prior studies have primarily demonstrated the utility of TDA in distinct disease comparisons,

often overlooking complex phenotypic variations such as cancer stage and treatment response [22, 93, 96]. Its effectiveness in these more intricate clinical predictions remains underexplored. Addressing these harder prediction tasks can establish topological descriptors as a robust approach for data representation and feature selection in machine learning models for oncology.

2. **Optimizing Data Representation and Feature Selection in TDA** While TDA inherently provides a powerful framework for data representation and feature selection, there is a notable lack of research comparing different metric and measure spaces to evaluate their effectiveness. The construction of simplicial complexes has predominantly relied on correlation-based methods, yet alternative distance measures remain largely unexplored, despite their potential to generate more informative and meaningful topological representations.

To overcome the two challenges in clinical prediction, this thesis specifically focuses on:

1. **Predicting Complex Phenotype Variations Across Diverse Datasets.** Despite the promise of TDA, prior research has primarily focused on simpler phenotype predictions, leaving more complex clinical outcomes underexplored. This thesis addresses this gap by leveraging topological descriptors to predict cancer type, stage and treatment response. Each of these clinical outcomes presents increasing levels of complexity, requiring progressively deeper insights into tumor biology. By systematically addressing increasingly challenging prediction tasks, we aim to establish TDA as a powerful tool for modeling complex phenotypic variations for improving machine learning models.

2. **Optimizing Measures for Enhanced Data Representation and Feature Selection:** This thesis aims to improve data representation and feature selection in downstream machine learning tasks by evaluating alternative metric and measures. Specifically, we compare Pearson Correlation, Distance Correlation, wTO using Pearson-based adjacency, and wTO with Distance-based adjacency to determine their impact on downstream model performance. This directly addresses the research aim 1.2.4.2: *Compare different co-expression measures in constructing the simplicial complex and assess their influence on model performance.* By evaluating the choice of co-expression measure used in the simplicial complex construction, this aim seeks to understand the impact of the measure on the predictive performance of downstream machine learning models.

By addressing these limitations, this thesis establishes TDA as a powerful complement to statistical and machine learning approaches in predicting clinical outcomes.

Ultimately, this work underscores the transformative potential of topology-driven analytics in advancing precision medicine. By bridging the gap between theory and practical application, this research contributes to the development of robust methodologies for biomarker discovery and clinical outcome prediction, with the goal of improving patient care and validating new topological based methodologies for our understanding of cancer biology.

Chapter 4

Materials and Methods

This study applies persistent homology to biomarker discovery and clinical outcome prediction using gene expression data. By incorporating TDA into the clinical workflow, we aim to uncover meaningful gene interactions and enhance predictive accuracy. This section outlines the methods used to achieve these objectives, detailing data preprocessing, co-expression measures, persistent homology, machine learning integration and the visualization tool built for WGTDA.

4.1 Overview

4.1.1 Clinical Outcome Prediction

The clinical prediction pipeline begins with rigorous preprocessing and gene set pre-selection, designed to ensure data quality and focus on biologically relevant features. This includes sample and gene filtering, outlier detection and adjustment, and Differential Expression Analysis through *DESeq2*. This preprocessing and pre-selection step aims to focus on relevant molecular features and reduce the computational expense associated with computing persistent homology.

The selected gene sets were encoded into co-expression matrices using four distinct measures. Namely, Pearson Correlation, Distance Correlation, Weighted Topological Overlapping (wTO) with Pearson-based adjacency and wTO using Distance-based adjacency. These co-expression matrices were transformed into patient-specific weights by integrating each patient's unique gene expression profile through an adjustment formula. This adjustment term incorporates the expression levels of each patient, ensuring that the computed weights reflect the relative co-expression strength between genes. The patient-weights were transformed into patient-specific VR complexes, enabling the extraction of shapes and structures embedded in each patient. Persistent homology was applied to these complexes to uncover topological features such as connected components (β_0), loops (β_1), and voids (β_2). To integrate these features into machine learning models, persistent landscapes were generated. This vectorization procedure transforms topological information into a format compatible with predictive

modeling.

The study evaluates the impact of TDA-derived features by comparing the performance of machine learning models with and without these features. The machine learning models that included the gene expression data itself (and not the TDA-derived landscapes) include SVM, Neural Networks, and Light Gradient Boost Machines (LightGBM). Only one machine learning model included persistent landscapes and gene expression data which was Random Forest. We predicted on tasks like cancer types, cancer staging and treatment response. Additionally, we assess the influence of different co-expression measures on predictive accuracy, providing insights into the optimal representation of gene interactions for clinical prediction tasks. The summarized framework is depicted in Figure 4.1.

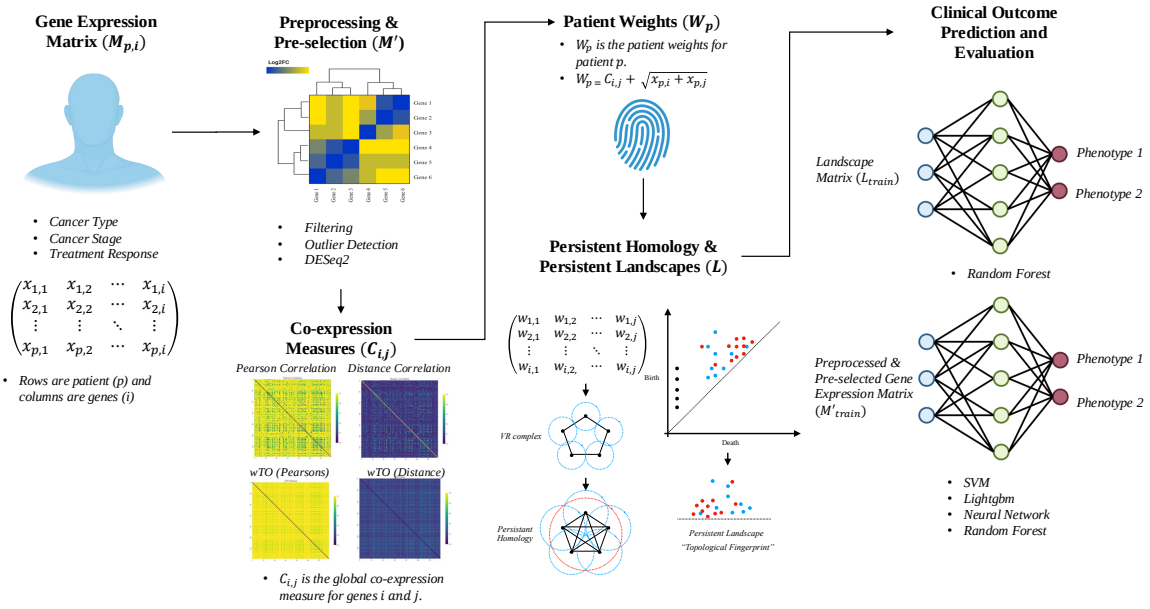


Fig. 4.1 Clinical Outcome Prediction Framework. The framework illustrates the methodology for integrating persistence landscapes with machine learning models to predict clinical outcomes. Starting with the gene expression matrix $M_{p,i}$ the data undergoes preprocessing and gene-set pre-selection $M'_{p,i}$. Four co-expression measures are applied $M'_{p,i}$ and the co-expression matrices ($C_{i,j}$) are used to compute weights for every patient (W_p) through a adjustment term ($\sqrt{x_{p,i} + x_{p,j}}$). This adjustment term incorporates the expression levels of each patient, ensuring the weights reflect relative co-expression strength between genes. Persistent homology is then used to extract topological features, which are transformed into persistence landscapes and subsequently used for classification. Additionally, the preprocessed gene expression matrix ($M'_{p,i}$) is also used for classification, enabling a comparison of performance between topological descriptors and traditional gene expression data.

4.1.2 Biomarker Discovery

Similar to the clinical outcome prediction pipeline, biomarker discovery began with preprocessing and gene set pre-selection, including sample and gene filtering, outlier adjustment, and differential expression analysis, to prioritize relevant features and reduce computational demands.

Following this, the pre-selected gene sets were encoded into a single global co-expression matrix using the best-performing co-expression measure ($C'_{i,j}$) identified during clinical outcome prediction. This matrix captured global gene interactions and served as the foundation for constructing the VR complex. In this framework, genes were represented as 0-simplices (nodes), while gene interactions formed higher-dimensional simplices. Persistent homology was applied to these complexes to extract topological features, focusing on β_1 (loops) and β_2 (voids), which highlight structural patterns within the co-expression matrix. To ensure robustness, the analysis retained only the top 3% of persistent features for β_1 and β_2, filtering out noise and emphasizing relevant topological structures.

To translate these abstract topological features into actionable insights, we developed a dual visualization framework that bridges the gap between topological representations and biological interpretation. The static visualization tool provided a high-resolution, global overview of gene interaction networks, scaling edges based on connectivity to highlight critical gene interactions. Complementing this, the web-based interactive visualization tool offered enhanced exploratory capabilities, dynamically scaling nodes and edges based on their degree of connectivity. This interactive platform allowed users to zoom, pan, and engage with the network, enabling the identification of key genes and interactions[1].

The final step in this pipeline was the validation of identified biomarkers through external literature, confirming their biological significance and clinical relevance. This thorough validation process demonstrated the robustness of the WGTDA framework and its ability to uncover meaningful gene signatures that can be leveraged as proposed prognostic biomarkers. Figure 4.2 illustrates the biomarker discovery pipeline.

[1]The web-based visualization tool is found here: https://nnyase.github.io/MSc-Thesis/

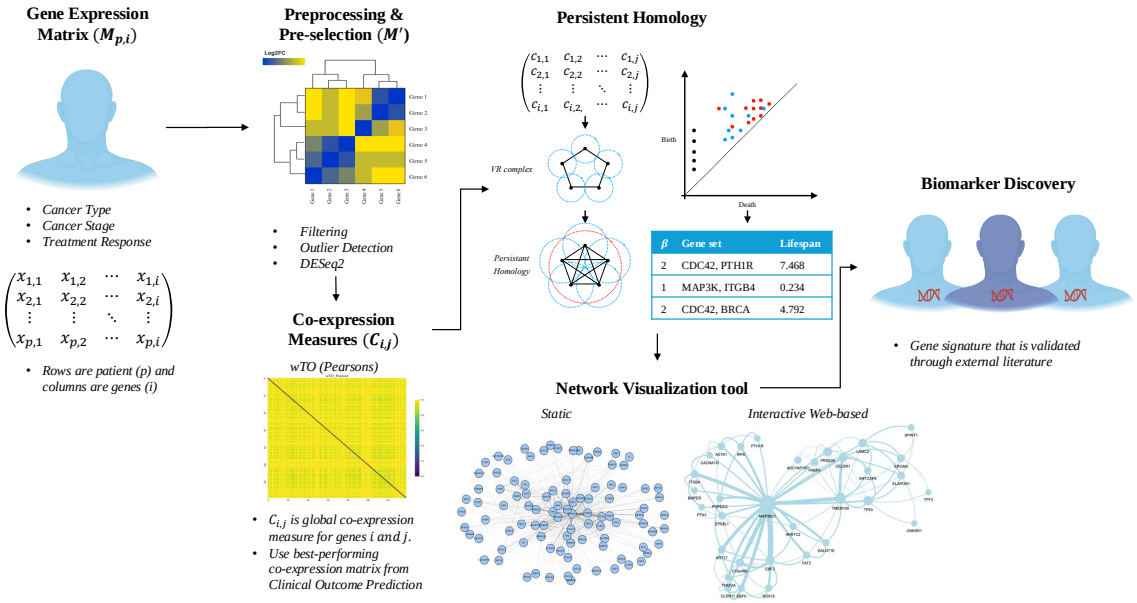


Fig. 4.2 Biomarker Discovery Framework. This framework illustrates the methodology for identifying prognostic biomarkers using WGTDA. Starting with the gene expression matrix ($M_{p,i}$), the data undergoes preprocessing and gene-set pre-selection to form $M'_{p,i}$. The best-performing co-expression measure ($C'_{i,j}$) from the clinical prediction tasks is selected as the co-expression matrix to construct our VR complex. Following this, we generate our topological features using persistent homology. The topological features are visualized using a dual network visualization tool, to explore gene-gene interactions and identify prognostic biomarkers. The proposed biomarkers, validated through external literature to establish clinical relevancy.

4.2 Data Acquisition

To address the research objectives of clinical outcome prediction and biomarker discovery, we curated data from prominent public repositories, primarily TCGA and CPTAC. Our focus was on protein-coding genes exclusively, as they represent the functional elements of the genome. The datasets were systematically organized into three tasks, each representing an increasing level of biological and clinical complexity, mirroring the progressive challenges associated with understanding and managing cancer.

4.2.1 Task 1: Cancer Type

The first task focuses on distinguishing between distinct cancer types, a foundational step in cancer diagnosis. For this, gene expression data were obtained from TCGA for the following cancers:

- **TCGA-SARC**: Sarcoma ($N = 259$) [2]

[2]TCGA-SARC can be accessed at GDC TCGA-SARC project page

- **TCGA-ESCA**: Esophageal Carcinoma ($N = 184$) [3]
- **TCGA-PCPG**: Pheochromocytoma and Paraganglioma ($N = 179$) [4]

Since these cancers represent distinct disease types, we hypothesize that classification will be relatively straightforward. However, this task serves as a crucial baseline to evaluate the effectiveness of topological features in capturing gene expression patterns before applying them to more complex phenotypes such as cancer staging and response to treatment.

4.2.2 Task 2: Cancer Staging

The second task introduces increased biological complexity by addressing cancer staging, which reflects tumor progression and metastasis. The dataset for this task is TCGA-HNSC (Head and Neck Squamous Cell Carcinoma)[5], comprising of 487 samples stratified into clinical stages I, II, III, and IV:

- **Stage I:** $N = 33$
- **Stage II:** $N = 148$
- **Stage III:** $N = 132$
- **Stage IV:** $N = 173$

Accurately predicting tumor stage and identifying stage-specific biomarkers is inherently more difficult than classifying cancer types. This task demands the detection of subtle, yet critical, patterns in gene expression data that are associated with tumor progression.

4.2.3 Task 3: Treatment Response

The third task addresses one of the most clinically challenging problems—predicting treatment response. For this task, we utilized data from CPTAC-3 radiotherapy response data across multiple cancer types and anatomical regions[6]. The samples are categorized into two groups:

- **Sensitive/Responsive:** $N = 347$
- **Resistant/Non-responsive:** $N = 106$

The data is further stratified by anatomical region:

- Brain ($N = 107$)

[3]TCGA-ESCA can be accessed at GDC TCGA-ESCA project page
[4]TCGA-PCPG can be accessed at GDC TCGA-PCPG project page
[5]TCGA-HNSC can be accessed at GDC TCGA-HNSC project page
[6]CPTAC data portal can be accessed at CPTAC data portal

- Uterus ($N = 120$)
- Lung ($N = 43$)
- Pancreas ($N = 23$)
- Kidney ($N = 8$)
- Other Regions ($N = 67$)

This dataset represents the highest level of complexity, as treatment resistance is influenced by multiple biological and clinical factors. By integrating radiotherapy response data, we aim to identify biomarkers and predict response to treatment, enabling more personalized therapeutic strategies. Additionally, we intend to leverage CPTAC data, recognizing that CPTAC and TCGA datasets differ in materials, measurements, and methodologies. Utilizing diverse data sources ensures robustness and enhances the generalizability of topological features and the application of WGTDA.

4.3 Preprocessing and Pre-selection

Data preprocessing and gene set pre-selection are essential initial steps in all three tasks, ensuring the raw RNA-Seq data is refined into a high-quality, and filtered dataset for topological analysis. Table 4.1 provides a summary of the preprocessing and pre-selection tasks[7].

4.3.1 Data Preprocessing

Preprocessing begins with systematically filtering of low-quality samples and genes to retain biologically relevant data. Samples with low counts across most genes were removed, more specifically those expressing fewer than 500 genes with counts ≤ 10. Similarly, genes with expression counts ≤ 10 in at-least 80% of samples were excluded. This stringent filtering approach aimed to eliminate noise and preserved biologically relevant data.

To address the potential impact of outliers, Cooks' distances was employed to flag and down-weight outlier genes. Cook's distance is a statistical estimate for detecting influential data points in least-square regression analyses by measuring their coefficients when removed [130]. The formula for Cook's distance is given by:

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{y}_j - \widehat{y}_{j(i)}\right)^2}{p * MSE} \tag{4.1}$$

[7]Code for the data preprocessing and gene set pre-selection task can be accessed at github.com/nnyase/MSc-Thesis/Data_Preprocessing.R

where:

- $\widehat{y}_j$ represents the predicted value for the j-th sample,
- $\widehat{y}_{j(i)}$ represents the predicted value when the i-th gene is excluded,
- p is the number of predictors
- MSE is the mean squared error (MSE) of the regression model,
- n is the total number of observations,
- and D_i is Cook's distance for i-th gene.

Genes exceeding the predefined Cook's distance threshold were adjusted, ensuring that the expression values were free from inconsistencies that could skew the analysis.

Finally, RNA-Seq data were normalized using Fragments Per Kilobase Million (FPKM). We employed FPKM as the measure accounts for both the length of a gene and the number of reads mapped to it, providing a normalized measure that allows for accurate comparisons across different genes and experimental conditions. This was done to mitigating biases associated with variations in gene length, enabling us to focus on relative expression levels rather than absolute read counts [131].

The approaches of filtering, outlier adjustment and normalization aimed to enhanced data quality and reduced the computational load, an important consideration for the resource-intensive nature of topology-based analyses.

4.3.2 Gene Set Pre-selection

Following preprocessing, a gene selection strategy was employed to focus on biologically meaningful processes while minimizing the computational burden associated with identifying topological features. To achieve this, the number of protein-coding genes was reduced through a two-stage gene set pre-selection process:

1. **Systematic Filtering:** As outlined in Section 4.3.1, low-quality genes and samples were removed.
2. **Differential Gene Expression Analysis:** This step identified up-regulated and down-regulated genes across different conditions.

The resulting gene set was used to construct the $M_{i,j}$ matrix, which serves as input to the subsequent construction of the simplicial complex.

4.3.2.1 Differential Gene Expression Analysis

Differential gene expression analysis was performed to identify significantly up-regulated and down-regulated genes, enabling the comparison of biologically distinct conditions across datasets while reducing dimensionality for downstream TDA [132]. This analysis was implemented in R using the *DESeq2* package (v1.46.0), a robust tool for moderated RNA-Seq data analysis [133]

The first step in the *DESeq2* workflow began with the normalization of raw count data. Size factors (s_j) were estimated for each gene j to account for the differences in library size and composition, ensuring that count values were comparable across samples. The mean count $\mu_{i,j}$ for each gene j in sample i was calculated as follows:

$$\mu_{i,j} = s_i \cdot q_{i,j} \tag{4.2}$$

where:

- $q_{i,j}$ represents the expression strength of gene j in sample i
- and s_i is the size factor that corrects the mean for technical variability.

Following normalization, gene-wise dispersion ϕ_j were estimated to capture the variability between biological replicates within the same group. Using data from seven replicates per group, it was assumed that genes with similar expression levels would exhibit similar dispersion values.

Moreover, to identify significant Differentially Expressed Genes (DEGs), a generalized linear model (GLM) was fitted to the normalized counts for each gene j, assuming a negative binomial (NB) distribution.

$$y_{i,j} \sim NB(\mu_{i,j}, \phi^j) \tag{4.3}$$

Here:

- $y_{i,j}$ is the observed count for gene j in the sample i,
- $\mu_{i,j}$ is the mean count and σ^2 is the dispersion parameter.

Differential expression was assessed using the likelihood ratio test (LRT), which evaluates all levels of a factor simultaneously rather than relying on pairwise comparisons. To control for false discovery rate, p-values were adjusted using the Benjamini-Hochberg (BH) method, selecting genes with an adjusted p-value ($padj$) ≤ 0.05 as significant DEGs.

Furthermore, to construct $M_{i,j}$ matrix for downstream topological analysis, genes with a log_2 fold change (log_2FC) greater than 1.5 or less than -1.5 were selected. However, for Task 1 (cancer type) a more stringent threshold of $log2FC$ of 4 was used to capture distinct tissue-specific gene expression differences. A lower threshold for this task would have included too many genes, thus increasing computational complexity and potentially diluting biological relevance.

This gene set pre-selection workflow provided a high-quality gene set and narrowed the vast genomic search space for downstream topological analysis. The resulting $M_{i,j}$ matrix was used for generating the co-expression matrix which is important for identifying biomarkers and for computing the patient's topological fingerprint.

Step	**Description**	**Method/Tool**	**Threshold**	**Outcome**
1	Sample filtering	Filtering	Samples with $\geq$ 500 expressed genes	High-quality samples retained
2	Gene filtering	Filtering	Genes expressed in $\geq 80\%$ of samples	High-quality genes retained
3	Detection and adjustment of outliers	Cook's distance	Cook's distance > threshold	Outlier genes down-weighted
4	Differential expression analysis	DESeq2	$p_{adj} \leq 0.05$ and $log_2FC \geq \|1.5\|$ (or $\|4\|$ in Task 1)	Significant DEGs identified
5	Normalization of RNA-Seq data	FPKM formula	Normalization of gene counts	Normalized RNA-Seq data

Table 4.1 Summary of Preprocessing and Gene Pre-selection Tasks

4.4 Co-expression Measures

Co-expresssion measures are methods to define networks among genes and usually fall into two categories correlation coefficients and mutual information measures [134]. They are used to measure relations between genes, compare conditions and characterize gene to gene relations to reveal insight on a biological condition [134, 135]. The research aim outlined in Section 1.2.4.2 seeks to evaluate and compare different co-expression measures for constructing informative simplicial complexes. This step is critical as accurately representing the relationships and interactions between genes will consequentially revealing meaningful topological features for downstream ML models and biomarker discovery [3, 22].

To achieve this, four distinct co-expression measures were examined:

1. **Pearsons Correlation** - Captures linear relationships.
2. **Distance Correlation** - Captures linear and non-linear dependencies.
3. **wTO using Pearson-based adjacency** - Incorporates shared network neighbors to assess gene connectivity through Pearson Correlation.
4. **wTO using Distance-based adjacency** - Incorporates shared network neighbors to assess gene connectivity through Distance Correlation.

These measures were used to construct the simplicial complex for the clinical outcomes pipeline, directly influencing the extracted topological features and overall model performance. The best-performing co-expression measure was then applied to biomarker identification using WGTDA, ensuring robustness and biological relevance.

4.4.1 Clinical Outcome Prediction

Before computing the co-expression measure for the clinical prediction pipeline, we randomly split the pre-selected gene expression matrix ($M_{i,j}$) into training (M_{train}) and testing (M_{test}) sets at 80% and 20% respectively. This random split minimizes the risk of unintended biases influencing the construction of the simplicial complex or the subsequent model's performance. To ensure consistency and reproducibility, a fixed random seed was used during the splitting process. The next step involves computing the different co-expression measures to construct the simplicial complex.

4.4.1.1 Pearson's Correlation

Pearson's Correlation ($r_{i,j}$) quantifies the strength and direction of a linear relationship between two arbitrary genes, x and y [136]. It is calculated using the formula:

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{k,i} - \bar{x}_i)(y_{k,j} - \bar{y}_j)}{\sqrt{\sum_{k=1}^{n}(x_{k,i} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{n}(y_{k,j} - \bar{y}_j)^2}} \tag{4.4}$$

where:

- $x_{k,i}$ and $y_{k,j}$ are the k-th observations of genes i and j,
- $\bar{x}_i$ and $\bar{y}_j$ are the mean expression levels of genes i and j,
- r_{ij} is the resulting Pearson correlation coefficient for genes i and j.

Pearson's Correlation was utilized to identify linear relationships within the training set (M_{train}). The computed coefficients are assembled into a symmetric correlation matrix $R_{i,j}$, where each entry $r_{i,j}$ represents the Pearson coefficient for genes x_i and x_j. Finally, the calculated $R_{i,j}$ correlation matrix was used to weight patients for the construction of the simplicial complex. The Pearson's correlation coefficients were calculated using the `scipy.stats.pearsonr` function in Python [137].

4.4.1.2 Distance Correlation

Distance Correlation extends on Pearson Correlation by addressing two key limitations. Firstly, it is applicable to random variables X and Y in arbitrary dimensions, making it more versatile than Pearson, which is limited to scalar variables[8]. Secondly, Distance Correlation captures a broader notion of dependence, with $R(X,Y) = 0$ indicating that X and Y are statistically independent, in contrast to Pearson Correlation, which can only detect linear relationships [138]. The Distance Correlation between two genes i and j is defined as:

$$dCor_{i,j} = \frac{dCov(x_i, y_j)}{\sqrt{dVar(x_i)} \cdot dVar(y_i)} \tag{4.5}$$

where:

- Distance covariance ($dCov(x_i, y_i)$) measures the dependency between genes x_i and y_j capturing both linear and nonlinear associations.
- Distance variance ($dVar$) measure of the variability within a single gene.

To compute the Distance Correlation matrix ($D_{i,j}$), Distance coefficients ($dCor_{i,j}$) were calculated for each pairs of genes x_i and y_j in the training set M_{train}. The resulting symmetric matrix ($D_{i,j}$) was used to construct the simplicial complex, where $dCor_{i,j}$ defined the edge's weights connecting the genes/simplices. Distance correlation coefficients $dCor_{i,j}$ was computed using `dcor.distance_correlation` in Python [139].

4.4.1.3 Topological Overlapping Measures $TOMs$

Topological Overlapping Measures ($TOMs$) quantify the similarity between genes by incorporating both their pairwise connections and shared neighbors within the network [140]. Unlike pairwise correlation methods such as Pearson's and Distance Correlation, which solely focuses on pairwise relationships, $TOMs$ leverages the underlying network structure to capture the broader connectivity patterns between genes. This makes $TOMs$ particularly effective for improved fidelity gene co-expression networks [140–142].

[8] In this thesis, random variables X and Y correspond to genes i and j.

4.4.1.4 Weighted Topological Overlap (wTO)

Weighted Topological Overlap (wTO) extends the concept of TOM by incorporating weights into the adjacency matrix, allowing for a more nuanced representation of gene connectivity. Unlike unweighted adjacency matrices $a_{i,j}$, where connections between gene pairs are binary (1 for connected, 0 if otherwise), wTO employs a weighted adjacency matrix (A_{ij}) to encode the strength of the pairwise connections with $A_{ij} \in [0,1]$. The weighted adjacency matrices can be built using any correlation or mutual information measure. In this study, we use Pearson and Distance Correlation to construct the weighted adjacency matrices for wTO.

The weighted adjacency matrix (A_{ij}) is computed as follows:

$$A_{i,j} = \begin{cases} |w_{i,j}|^{\beta} & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases} \tag{4.6}$$

where:

- $w_{i,j}$ is the pairwise correlation coefficient (e.g., Pearson's correlation or Distance correlation) between genes i and j,
- β is the soft threshold parameter that amplifies strong connections while suppressing weaker ones,
- $A_{i,j}$ is the resulting adjacency value representing the strength of the relationship between genes i and j.

Following the construction of the adjacency matrix, the $wTO_{i,j}$ calculated to quantify the similarity between two genes i and j, incorporating both their direct connection and shared neighbors. The formula was adapted from Zhang et al. [141] is:.

$$wTO_{i,j} = \frac{\sum_{k \neq i,j} A_{i,k} \cdot A_{j,k} + A_{i,j}}{min(K_i, K_j) + 1 - |A_{i,j}|} \tag{4.7}$$

Here:

- $A_{i,j}$ is the adjacency value for genes i and j,
- k represents the common neighbors in the network,
- $\sum_{k \neq i,j} A_{i,k} \cdot A_{j,k}$ refers to the shared connectivity between genes i and j , computed as the dot product of their connections to common neighbors k,

- $min(K_i, K_j)$ Compute the minimum of the weighted degree (sum of absolute connections) between genes i and j, which is adjusted by 1 - $A_{i,j}$ as a direct correction.

To compute the $wTO_{i,j}$ matrix, we began by constructing adjacency matrix $A_{i,j}$ using both Pearson's and Distance correlation as pairwise gene relationships in M_{train}. The shared neighborhood between two genes i and j is computed as the dot product of row and column vectors in the adjacency matrix $A_{i,j}$. This is normalized by the minimum value of the weighted degree K_i and K_j ensuring relative measure of connectivity. The implementation of wTO was developed in Python and is accessible through `ibm.wgtda.compute_wto_matrix`. The function was adapted to include Pearson and Distance correlation as pairwise adjacencies for this thesis. The resulting symmetric matrix $wTO_{i,j}$ was used to extract patient weights to generate a topological fingerprint for each patient.

For consistency throughout the work, we refer to these co-expression matrices ($R_{i,j}$, $dCor_{i,j}$, $wTO_{i,j}$) collectively as $C_{i,j}$. We also refer to the best-performing co-expression measure as $C'_{i,j}$

$$C_{i,j} = \begin{cases} R_{i,j} & \text{if derived from Pearson's correlation,} \\ dCor_{i,j} & \text{if derived from Distance correlation,} \\ wTO_{i,j} & \text{if derived from wTO.} \end{cases}$$

Ultimately, the pre-processed gene expression matrix (M'), containing normalized expression values for differentially expressed genes (DEGs), was transformed into co-expression matrices ($C_{i,j}$). These matrices provide a global representation of gene-gene relationships, enabling the construction of informative simplicial structures essential for downstream clinical outcome prediction.

4.4.2 Biomarker Discovery

The biomarker discovery approach focuses on generating a single co-expression matrix for biomarker identification, following the methodology used for clinical outcome prediction outlined in Section 4.4.1. However, two key modifications were made to tailor this process specifically for identifying prognostic biomarkers: (1) Leveraging the best-performing co-expression measure ($C'_{i,j}$) from the clinical outcome models to ensure optimal biomarker selection, and (2) Utilizing the entire dataset without splitting to maximize statistical power and capture a more comprehensive biological signal.

1. **Leveraging the Best-Performing Co-expression Measure ($C'_{i,j}$)**

   Instead of independently evaluating multiple co-expression measures, biomarker discovery utilizes the best-performing co-expression measure identified during the clinical prediction phase. This ensures that the selected co-expression metric is both optimal and biologically meaningful, leveraging insights from its predictive performance.

2. **Utilizing the Entire Dataset Without Splitting**

   To maximize the statistical power and ensure that all data contributed to the construction of the co-expression network, the pre-selected gene expression matrix M' was used without splitting it into training and testing sets. Unlike the clinical prediction framework, where the data was partitioned to evaluate model generalizability, the entire dataset M' was designated as M_{train} for biomarker discovery. In this case, M_train represents the full dataset of M'. Since the primary objective was biomarker discovery rather than predictive validation, a separate testing set was not required at this stage.

By integrating the optimal co-expression measure in clinical prediction and leveraging the entire dataset, the biomarker discovery process ensures that relationships between genes are captured comprehensively. The transformation of pre-processed gene expression matrix M' into $C'_{i,j}$ for biomarker identification was implemented through the WGTDA Python package.

4.5 Patient Weights for Prediction

For clinical outcome prediction, we tailor the global co-expression matrix C to individual patient profiles, by introducing inter-gene patient weights that adjust the gene-gene relationships based on each patient's unique gene expression levels. This adjustment transforms the population-wide co-expression structure $C_{i,j}$ into a patient-specific weighted matrix W_p through a scaling parameter. The resulting patient-specific distance matrix serves as the input for constructing the simplicial complex.

By integrating global co-expression patterns with individual expression variability, this approach aims to create personalized topological fingerprints that capture both shared co-expression patterns and patient-specific nuances. These fingerprints offer a unique representation of gene connectivity for each patient, enhancing the ability to identify distinct patient profiles for clinical outcome prediction.

The patient-specific weighted matrix (W_p) is derived by adjusting the global matrix (C) to reflect the unique expression profile of each patient p. For every gene pair i and j, the relationship is modified using the following formula:

$$W_p[i,j] = C_{i,j} + \frac{\sqrt{x_{p,i} + x_{p,j}}}{10}, \tag{4.8}$$

Where:

- C represents the gene co-expression matrix,
- $x_{p,i}$ and $x_{p,j}$ are the expression values of gene i and gene j for patient p,

- W_p is the weighted matrix for patient p.

The adjustment term ($\sqrt{x_{p,i}+x_{p,j}}$) introduces a controlled, nonlinear scaling effect emphasizing strong relationships for highly expressed genes while minimizing the influence of low-expression relationships compared to the global population. By integrating patient-specific variability into the global co-expression structure, the resulting symmetric weighted matrix W_p balances shared population-wide trends with individual expression differences.

An intuitive analogy is that the global co-expression matrix C is a global roadmap for gene relationships across the population. It provides a baseline structure that shows which genes tend to co-express or interact in the broader dataset. This can be seen as a general navigation map that works for everyone but doesn't account for individual preferences or unique routes. The adjustment term ($\sqrt{x_{p,i}+x_{p,j}}$) personalizes this map based on how strongly each gene is expressed for a given patient p. This is like tailoring the navigation map to include personal preferences, prioritizing frequently traveled roads (high gene expression) while downplaying less relevant ones (low gene expression). The final matrix W_p is now customized for each patient, it reflects the general trends across the population but also the unique expression levels of that patient, thus prioritizing the context the patients are situated in but also their unique differences. The formula customizes the global gene-gene relationships to account for what is most relevant for a specific patient, ensuring a personalized simplicial complex which will be utilized for classification.

4.6 Simplicial Complex Construction

In this study, VR complexes are used as the foundation for analyzing gene expression data. The VR complex was chosen over alternatives such as the Čech complex due to its lower memory requirements and computational efficiency [23, 43]. Persistent homology is applied to these complexes to extract topological features that capture both local and global homology groups within the data. These features are then used for clinical outcome prediction and biomarker discovery, with tailored construction methods reflecting the objectives of each task.

For clinical outcome prediction, the VR complex is generated from patient-specific weighted matrices (W_p), which adjust global co-expression patterns to reflect individual gene expression variability. In contrast, for biomarker discovery, WGTDA constructs the VR complex using the global gene co-expression matrix (C) for each task, capturing shared relationships across genes at the population level.

4.6.1 VR Complex for Clinical Outcome Prediction

For clinical outcome prediction, we generated p patient-specific VR complexes for each co-expression measure, where p is the total number of patients. Each patient p was assigned a unique weighted matrix

W_p, derived by integrating their individual gene expression values with a global population-wide co-expression matrix. We denote the VR complex generated for each patient as VR_p with p representing the patient. We implemented the construction using the *Gudhi* library (version 3.10.1), which is a library for computational topology and TDA [143].

To balance computational efficiency with accuracy, we employed sparse approximations during construction. Moreover, the maximum edge length was set to ∞ ensuring that no pairwise relationships were excluded during the initial filtration process. To further optimize the construction, edge collapse was applied. Edge collapse is a backwards reduction method which effectively collapses redundant edges, reducing the input flag filtration to its 1-skeleton, which consists of vertices and edges only. By simplifying the complex to its 1-skeleton, we minimized memory usage and computational overhead, enabling scalability for high-dimensional gene expression data [144].

Once the 1-skeleton was established, higher-dimensional simplices were expanded up to a dimension of 3. This step was critical for capturing higher-order relationships among the genes, such as loops (1D) and voids (2D and above), which are embedded within the gene expression data. The output of this process was a set of sparse VR complexes, one for each patient VR_p. Persistent homology was then applied to these complexes to extract topological features, which were subsequently used to generate topological fingerprints for downstream machine learning tasks.

4.6.2 VR Complex for Biomarker Discovery

In contrast to clinical outcome prediction, the VR complex for biomarker discovery was constructed using global co-expression matrices C using WGTDA rather than patient-specific weighted matrices. This approach enables the transformation of gene expression profiles across all patients into a single, unified topological object, capturing global patterns of gene-gene interactions at the disease level. This VR complex that captures global expression patterns is denoted as VR_t, where t is referencing the specific task (e.g., cancer type, cancer staging and treatment response). By focusing on the shared structure within the population, this method highlights consistent topological features that may serve as robust biomarkers for disease characterization. To generate the VR complex for biomarker discovery, we utilized the the *Gudhi* library (version 3.10.1).

To balance computational efficiency and accuracy, we employed the same steps as clinical outcome prediction, including sparse approximations, setting the maximum edge length to ∞ and applying edge collapse to reduce the VR complex to its 1-skeleton, thereby minimizing memory usage and enabling scalability for high-dimensional gene expression data [144].

The resulting global VR_t complex serves two critical purposes: (1) Allowing us to apply persistent homology to identify topological features that highlight prognostic biomarkers for each task. (2)

Providing the foundation for a custom visualization tool to explore and interpret the identified biomarkers, further enhancing the interpretability and biological relevance of the findings.

4.7 Persistent Homology

Persistent homology is a key step in extracting the topological features from the gene expression data, enabling the identification of meaningful structures that persist across multiple scales. This process extracts the birth and death of topological features, such as connected components (β_0), loops (β_1), and voids (β_2) in simplicial complexes. These features are subsequently used for clinical outcome prediction and biomarker discovery, with distinct approaches tailored to the goals of each task.

4.7.1 Persistent Homology for Clinical Outcome Prediction

For clinical outcome prediction, persistent homology was computed from the VR complexes for each patient (VR_p), capturing topological features at different dimensions: for β_0, β_1, and β_2. Higher-dimensional homology groups (β_3 and beyond) were excluded due to computational constraints.

4.7.1.1 Persistent Landscapes

To enable the integration of persistent homology into machine learning models, the persistence intervals for β_0, β_1, and β_2 were transformed into persistent landscapes. Persistent landscapes are a stable and robust vectorized representation of topological features as mentioned in Section 2.11.1. Persistent landscapes aggregate the birth-death intervals into functions over a fixed grid, enabling the capture of essential topological information while remaining resilient to noise and small perturbations in the data [51].

For this study, we set the number of landscapes to 2 per homology group and the resolution to 100, balancing computational efficiency with feature granularity. For each patient, landscapes were generated across the three key homology groups, resulting in three distinct landscapes $L0, L_1$ and L_2. These landscapes were then concatenated into a single feature matrix L_{train} and L_{test} for each patient. Following this, K-fold cross-validation was employed and the optimal model was selected of training using L_{train} and subsequently employed to predict outcomes on L_{test}.

4.7.2 Persistent Homology for Biomarker Discovery

For biomarker identification, persistent homology was applied to the global gene expression matrix C, focusing on β_1, and β_2 as these capture higher-order relationships such as loops and voids. β_0 features (connected components) were excluded due to their limited relevance for biomarker discovery, while β_3 and higher features were not calculated due to computational complexity [31].

The top 3% of features with the longest persistence (ε-range) were selected as biologically relevant signals, while short-lived features were treated as noise and removed, ensuring robust and meaningful interactions [31, 145]. This adopted approach in TDA prioritizes stable and significant patterns. The selected features were subsequently used to demonstrate the capabilities of the custom visualization tool and validated against external literature, highlighting the relevance and utility of WGTDA.

4.8 Predictive Modeling for Clinical Outcome Prediction

This section presents a systematic approach to clinical outcome prediction using both traditional machine learning algorithms (SVM, LightGBM, Neural Networks (NN), and Random Forest (RF)) and TDA-derived models. By comparing these two strategies, we aim to assess the predictive power of gene expression data alone versus topological fingerprints derived from TDA.

4.8.1 Overview of Approach

The models were trained using two distinct strategies:

1. **Baseline Models (Traditional ML):** SVM, LightGBM, Neural Networks, and Random Forest were trained directly on the preprocessed gene expression matrix (M_{train}) to evaluate the predictive capacity of the traditional ML models.
2. **TDA Models:** Only the Random Forest was specifically trained on persistent landscapes (L_{train}) derived from the VR complex constructed using four co-expression measures:
    - **Pearson's Correlation**
    - **Distance Correlation**
    - **wTO using Pearson-based adjacency**
    - **wTO using Distance-based adjacency**

By adopting these dual strategies, we obtain a broad evaluation of how traditional machine learning compares to a topological feature–driven approach for predicting clinical outcomes. Additionally, we also could assess the impact of different co-expression measures on the generation of topological features and their subsequent effect on model performance.

4.8.2 Cross Validation

To ensure a fair comparison between baseline and TDA models, we implemented 5-fold cross-validation (CV) on 80% of the training landscapes matrix (L_{train}) and the gene expression matrix (M_{train}) for hyperparameter tuning and model selection. The remaining 20% L_{test} and M_{test} was held out for final evaluation. This approach minimized potential biases and improved model reliability. We utilized *GridSearchCV* function from `scikit-learn` (version 1.6.0) [146].

4.8.3 Model Descriptions and Hyperparameters

Below, we outline the models and their corresponding hyperparameter grids explored during GridSearchCV.

4.8.3.1 Support Vector Machine (SVM)

SVM are powerful classifiers that find an optimal hyperplane to separate data points in high-dimensional space. In this study, SVMs were applied to the gene expression matrix M_{train} [147]. The following hyperparameter grids were explored:

- $\boldsymbol{C} \in$ [0.1, 1, 10, 100] (Regularization)
- **Kernel** $\in$ [linear, radial basis function, polynomial]
- **Gamma** $\in$ [scale, auto]

4.8.3.2 Light Gradient Boosting Machines (LightGBM)

LightGBM is a highly efficient gradient boosting decision tree (GBDT) that employs Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to accelerate training speed and reduce memory [148]. We utilized LightGBM to train M_{train} using the following hyperparameters:

- **n_estimators** $\in [50, 100, 200]$ (Number of boosting rounds)
- **max_depth** $\in [3, 5, 7]$ (Maximum depth of each tree)
- **learning_rate** $\in [0.01, 0.1, 0.2]$
- **subsample** $\in [0.8, 1.0]$ (Fraction of samples used per boosting round)

4.8.3.3 Neural Network

Neural Networks use interconnected layers of neurons with activation functions, enabling them to learn complex nonlinear relationships. The following hyperparameters were used for training on M_{train}:

- **Hidden Layer Size** $\in [(50,), (100,), (50, 50), (100, 50)]$
- **Activation Function** $\in$ [relu, tanh, logistic]
- **Alpha** $\alpha \in$: $[0.0001, 0.001, 0.01]$ (L2 regularization term)
- **Learning Rate** $\in$ [constant, adaptive]

4.8.3.4 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to improve accuracy and robustness [149]. It was trained on both topological fingerprints (L_{train}) and the gene expression matrix (M_{train}) using the following hyperparameters:

- **n_estimators** $\in [50, 100, 200]$ (Number of trees)
- **max_depth** $\in$ [None, $10, 20$] (Maximum depth of each tree)
- **min_samples_split** $\in [2, 5, 10]$ (Minimum number of samples required to split a node)
- **min_samples_leaf** $\in [1, 2, 4]$ (Minimum number of samples required at a leaf node)

These models were selected for their ability to capture both linear and complex non-linear relationships within the data, as well as their ensemble learning capabilities. This comprehensive approach leaves no room for doubt regarding the performance and generalizability of the findings. Furthermore, the systematic hyperparameter tuning process ensured that all models were optimized for clinical outcome prediction, enhancing the reliability and robustness of our findings.

4.8.4 Evaluation Metrics

To comprehensively evaluate the performance of the models, we utilized the following metrics: Accuracy, Mean Squared Error (MSE), Log Loss, F1-Score. These metrics provide a holistic view of model performance, capturing both predictive accuracy and the ability to evaluate imbalanced classes.

1. **Accuracy**: Accuracy measures the proportion of correct predictions out of the total number of predictions. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.9}$$

   where TP True Positives are correctly predicted positives, TN True Negatives are correctly predicted negatives), FP False Positives (incorrectly predicted positives), and FN False Negatives (incorrectly predicted negatives).

2. **Mean Squared Error (MSE)**: MSE evaluates the average squared difference between predicted and actual values. It is commonly used for regression tasks but can also be applied to classification probabilities:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4.10}$$

   where:

   - y_i: True value for the i-th sample,
   - $\hat{y}_i$: Predicted value for the i-th sample,

- n: Total number of samples.

3. **Log Loss**: Log Loss (or Cross-Entropy Loss) quantifies the performance of a classification model where the output is a probability between 0 and 1. It penalizes predictions that are far from the actual class labels:

$$\text{Log Loss} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)] \tag{4.11}$$

where:

- y_i: True label (0 or 1) for the i-th sample,
- $\hat{p}_i$: Predicted probability for the positive class,
- n: Total number of samples.

4. **F1-Score**: The F1-Score is the harmonic mean of precision and recall, balancing both false positives and false negatives. It is particularly useful for imbalanced datasets:

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.12}$$

where:

- Precision = $\frac{TP}{TP+FP}$,
- Recall = $\frac{TP}{TP+FN}$.

By combining robust model selection, hyperparameter tuning via grid search, and comprehensive evaluation, we ensured credible and reproducible predictive analysis. While SVM, LightGBM, and Neural Networks provided baseline results from raw data, incorporating persistent landscapes in the Random Forest model highlighted the added value of topological features for improving clinical outcome prediction.

4.9 Network Visualization Tool for Biomarker Discovery

Interpreting topological features can be challenging due to the ambiguity regarding which nodes formulate homology groups. Additionally, effective visualization of gene interactions is essential for understanding the complex relationships between genes, proteins, and other molecular entities. To address these challenges, we developed a network visualization tool that provides both static and interactive representations of topological features derived from WGTDA. This tool was built using *NetworkX* [150], *Matplotlib* [151], and *Pyvis* [152], enabling researchers to analyze and interpret topological features more effectively while enhancing their biological relevance.

4.9.1 Static Visualization Tool

The static visualization tool provides a clear, high-quality representation of gene interaction networks, enabling researchers to observe global patterns and identify key connections. The features of the static network are as follows:

1. **Network Construction:**
   (a) Genes are represented as nodes, while relationships between genes are depicted as edges.
   (b) Edge weights are computed based on the frequency of interactions between gene sets.
   (c) Thicker edges indicate stronger or more frequent interactions, providing visual emphasis on significant gene relationships.
2. **Network Features:**
   (a) Nodes are uniformly scaled for clarity, representing individual genes in the network.
   (b) A spring layout algorithm positions the nodes to minimize edge overlaps and optimize visualization clarity.

The static visualization is generated using Matplotlib, producing a visually interpretable graph that highlights the network's critical features and enables researchers to gain immediate insights into gene connectivity.

4.9.2 Interactive Visualization Tool

To complement the static visualization, we developed a web-based interactive tool for exploring persistent topological gene networks dynamically. This tool enhances interpretability by allowing users to interact directly with the network and explore its features in depth. The interactive visualization tool is accessible through a dedicated landing page: https://nnyase.github.io/MSc-Thesis/. Figure 4.3 shows the image of the dedicated landing page, where users can access the WGTDA web-based networks.

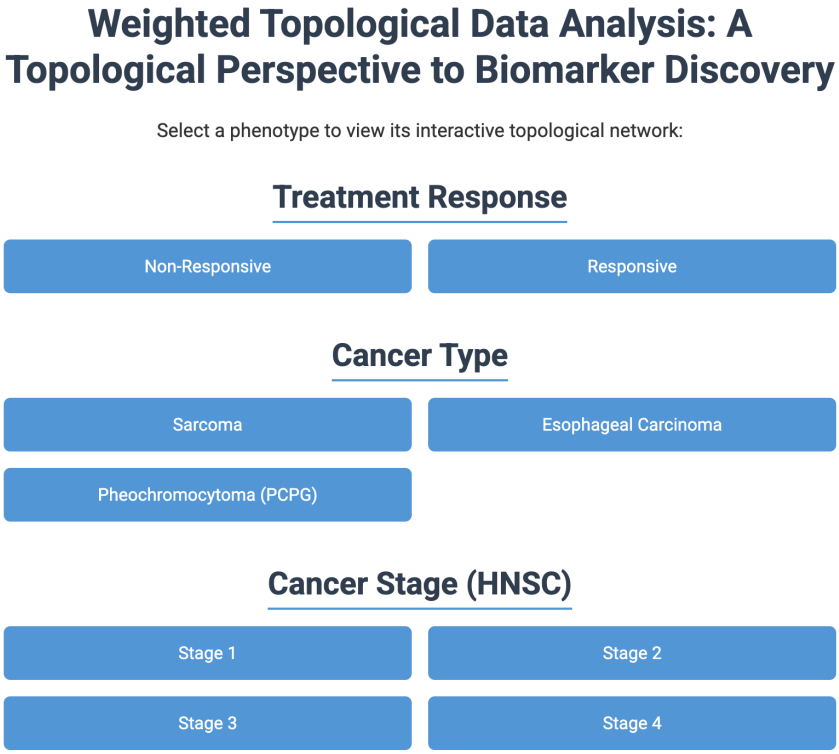


Fig. 4.3 Dedicated interactive landing page showing tasks for Cancer Type, and Cancer Stage (HNSC) and Treatment Response.

The features of the interactive web-based tool for WGTDA are as follows:

1. **Dynamic Node Scaling:** Nodes are scaled based on their degree of connectivity (i.e., the number of edges connected to a node). Therefore, genes with more interactions are visually prominent, allowing users to quickly identify key players in the network.

2. **Click Interactions:** Clicking over a node displays the gene name and the number of connections, offering immediate contextual information. Moreover, clicking on an edge reveals its weight (the number of interactions between the two connected genes), providing insights into the strength of specific relationships.

3. **Edge Visualization:** Edge widths dynamically adjust to reflect interaction strengths, with thicker edges denoting stronger relationships. This feature highlights critical gene interactions, making them easier to identify and interpret at a glance.

4. **Interactive Exploration:** Users can zoom, pan, and rearrange nodes to examine specific network regions or clusters in detail. A physics-based layout algorithm adds realism to the network, giving it a dynamic and engaging appearance. This functionality is particularly valuable for identifying clusters of related genes and exploring broader patterns across the network.

The interactive tool offers an intuitive platform for researchers to delve into the structure of gene interactions, bridging the gap between abstract topological features and meaningful biological insights.

4.10 Framework Summary

To provide clarity and enhance understanding, we summarize the experimental pipelines for both clinical outcome prediction and biomarker discovery. These pipelines outline the sequence of steps, from data preprocessing to the generation of topological features. Below, we present the structured workflows in algorithmic form for both tasks.

4.10.1 Clinical Outcome Prediction Framework

Algorithm 1 Clinical Outcome Prediction Pipeline

1: **Input:** Gene Expression Matrix X for cancer type, cancer stage and treatment response.
2: **Output:** Classification Results and Performance Metrics for the topological features and the preprocessed gene expression matrix.
3: **Step 1:** Preprocess the gene expression matrix X by filtering low quality samples and genes, outlier correction using Cook's distance and FKPM normalization.
4: **Step 2:** Perform differential expression analysis on the preprocessed gene expression data X' to obtain the gene set matrix M.
5: **Step 3:** Split the gene expression matrix M into training M_{train} and testing M_{test} splits at 80% and 20% respectively.
6: **Step 4:** Use gene expression matrix M to generate a global gene-gene co-expression matrix $C_{i,j}$ through Pearson's Correlation, Distance Correlation, wTO using Pearson's adjacency and wTO with Distance adjacency.
7: **Step 5:** Compute the patient weights W_p for each patient by extrapolating the global gene-gene co-expression matrix $C_{i,j}$ with the adjustment term $(\sqrt{x_{p,i} + x_{p,j}})$
8: **Step 6:** Construct the VR complex (VR_p) for each patient using the inter-gene patient weights W_p.
9: **Step 7:** Compute persistent homology for homology groups $\beta_0, \beta_1, \beta_2$ using the VR complex for each patient
10: **Step 8:** Generate the persistence landscapes (L_{train} and (L_{test}) for each patient-specific topological fingerprint.
11: **Step 9:** Train machine learning models (SVM, LightGBM, RF, NN) using cross validation for k-folds of M_{train} to establish baseline ($k = 5$).
12: **Step 10:** Train random forest model on the k-folds topological fingerprints (L_{train}) generated ($k = 5$).
13: **Step 11:** Extract best model from GridSearchCV parameters for each model.
14: **Step 12:** Perform model evaluation and report the performance metrics.

4.10.2 Biomarker Discovery Framework

Algorithm 2 WGTDA Pipeline

1: **Input:** Gene expression matrix X for cancer type, cancer stage and treatment response.
2: **Output:** Network Visualization of proposed prognostic biomarkers.
3: **Step 1:** Preprocess the gene expression matrix X by filtering low quality samples and genes, outlier correction using Cook's distance and FKPM normalization.
4: **Step 2:** Perform differential expression analysis on the preprocessed gene expression data X' to obtain the gene set matrix M.
5: **Step 3:** Use the full gene expression matrix M to generate a global gene-gene co-expression matrix $C_{i,j}$ using the best performing co-expression measure ($C'_{i,j}$) from the clinical predictions pipeline.
6: **Step 4:** Construct the VR complex (VR_t) for each task.
7: **Step 5:** Compute persistent homology for homology groups β_1, β_2 for each task.
8: **Step 6:** Extract topological features with associated genes formulating the feature.
9: **Step 7:** Create static and interactive visualization tool to have intuitive method of viewing topological features.
10: **Step 8:** Identify significant biomarkers based on persistent and network visualization.

Chapter 5

Results

In Chapter 5, the results of our study are presented in two distinct yet interwoven parts: clinical outcome prediction and biomarker discovery. The former focuses on leveraging patient-specific topological fingerprints to forecast outcomes in cancer type, staging, and treatment response. Whereas the latter aims to identifying key genes driving cancer biology through WGTDA, demonstrating a visualization tool built identifying key topological interactions.

We begin by detailing the data preprocessing and differential expression analysis, which were performed to refine and pre-select the gene expression matrix for downstream analysis. This is followed by a comprehensive exploratory data analysis (EDA), where dimensionality reduction techniques such as Principal Component Analysis (PCA) [153], t-Distributed Stochastic Neighbor Embedding (t-SNE) [154], and Uniform Manifold Approximation and Projection (UMAP) [155] were applied to visualize the gene expression profiles and uncover latent patterns to provide critical insights into the different tasks.

We then transition to the results of the clinical outcome prediction, where we conduct a comparative analysis between machine learning models trained on the raw gene expression data and those utilizing patient-specific topological landscapes.

Lastly, we provide results of the biomarker discovery, where the WGTDA framework was applied to identify key genes driving cancer biology. To further enhance the interpretability of our findings, we showcase a custom visualization tool developed to explore and interpret topological interactions. This tool enables researchers to seamlessly navigate gene connectivity networks and topological structures, providing an interactive means to identify and validate potential biomarkers.

5.1 Gene Set Pre-selection

To identify DEGs, *DESeq2* was applied to RNA-Seq data, revealing significantly up-regulated and down-regulated genes across each dataset. Table 5.1 summarizes the datasets, highlighting DEG patterns and the number of genes retained for downstream analysis. This pre-selection step streamlined subsequent analyses by focusing on biologically relevant genes, enhancing both interpretability and computational efficiency.

Task	Description	Dataset	Sample Size	No. Genes	Up-regulated Genes	Down-regulated Genes
1	Cancer Types	TCGA-SARC, TCGA-ESCA, TCGA-PCPG	SARC (261), ESCA (185), PCPG (179)	125	43	82
2	Cancer Staging	TCGA-HNSC	Stage I (33), Stage II (148), Stage III (132), Stage IV (173)	146	28	119
3	Treatment Response	CPTAC Brain, Uterus and Other Regions	Resistant (120), Sensitive (98)	129	88	41

Table 5.1 Overview of the datasets, and number of genes retained and the number of up-regulated and down-regulated genes. This table provides a clear summary of the differential gene expression patterns across the analyzed datasets.

5.2 Exploratory Data Analysis (EDA)

To gain a deeper understanding of the gene expression datasets and to uncover latent patterns, we performed a comprehensive EDA. Dimensionality reduction techniques such as PCA, t-SNE, and UMAP were implemented to visualize the complex relationships between genes and samples, and to contextualize the difficulties faced by the proposed TDA-based models and the raw machine learning models in learning boundaries for separating phenotypic classes for each dataset. Each technique was selected to highlight complementary aspects of the data: PCA captures global variance and linear relationships, t-SNE emphasizes local neighborhood structures, and UMAP balances global and local structure while preserving structural properties.

In our implementation, we used 2 principal components for PCA using the `scikit-learn` package. For t-SNE, we used 2 components with a perplexity of 30 (also via `scikit-learn`), and for UMAP we used 2 components, 15 neighbors, and a minimum distance of 0.1, using `umap-learn` package. By combining these methods, we seek to reinforce (or refute) observed

patterns of separability with a higher degree of confidence, provide intuitive visual insights and establish implications for downstream ML tasks.

5.2.1 Cancer Type EDA

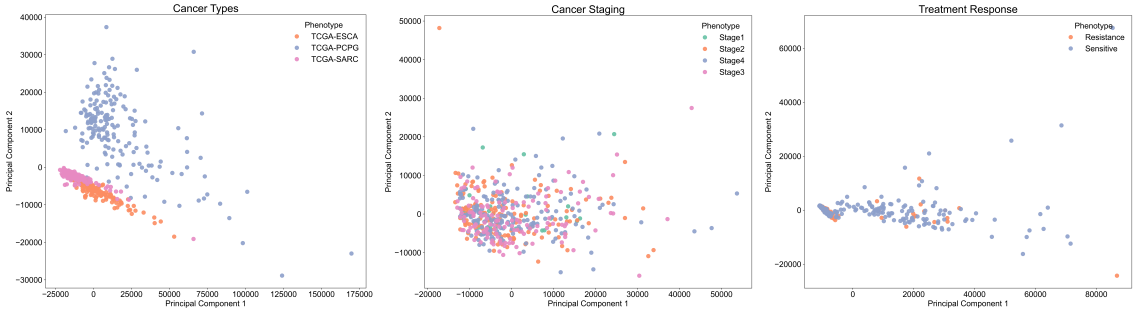
For cancer type tasks (SARC, ESCA, PCPG) the PCA plot shows clear separability between cancer types. ESCA and SARC form compact regions with slight overlap, while PCPG exhibits greater dispersion across the principal components, reflecting higher variance in its gene expression profiles. Both t-SNE and UMAP further reinforced these findings, showing well-defined and distinct clusters for each cancer type.

5.2.2 Cancer Staging EDA

In the cancer staging task (HNSC), PCA, t-SNE, and UMAP were applied to visualize the separation across stages I, II, III, and IV. All three methods revealed poor separability, and significant overlap between stages. The lack of clear boundaries suggests subtle differences in gene expression patterns across tumor stages, posing a challenge for downstream classification tasks.

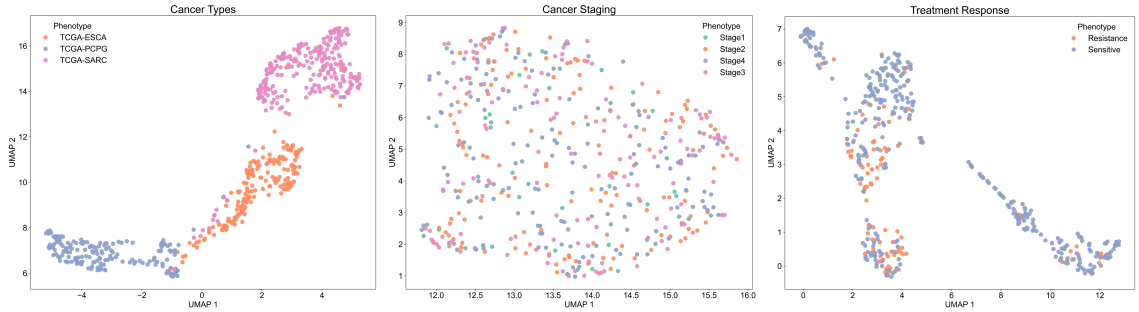
5.2.3 Treatment Response EDA

For the treatment response task, PCA, t-SNE, and UMAP were used to visualize separation between resistant and sensitive phenotypes. The dataset exhibited class imbalance, with sensitive samples outnumbering resistant ones. Additionally, all three techniques revealed poor separability, with substantial overlap between the two phenotypic groups, highlighting the complexity of accurately distinguishing treatment response based on gene expression profiles.



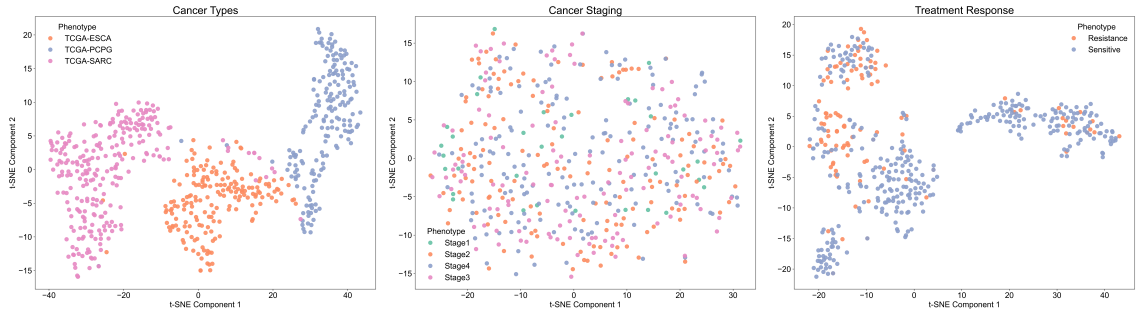
(a) Cancer Type (b) Cancer Staging (c) Treatment Response

Fig. 5.1 PCA for Cancer Type, Cancer Staging and Treatment Response



(a) Cancer Type (b) Cancer Staging (c) Treatment Response

Fig. 5.2 UMAP for Cancer Type, Cancer Staging and Treatment Response



(a) Cancer Type (b) Cancer Staging (c) Treatment Response

Fig. 5.3 TSNE for Cancer Type, Cancer Staging and Treatment Response

5.3 Clinical Outcome Prediction

This section presents the clinical outcome prediction results across the three tasks: cancer type classification, cancer staging, and treatment response. We focus on key evaluation metrics, including F1-score, training accuracy, and testing accuracy, to assess model performance and highlight the effectiveness of both traditional ML models and TDA-augmented models.

5.3.1 Cancer Type Classification

Table 5.2 presents the results for cancer type prediction[1]. The trend among conventional models trained on raw gene expression data-LightGBM, SVM, NN, and RF achieved high performance was consistently with F1-scores of 98–99% and testing accuracies ranging from 97.60-99.20%. These results highlight the robustness of these models in distinguishing between distinct cancer types using gene expression profiles.

[1]Classification results for cancer type prediction can be found here: Cancer Type Classification Code and Results

Moreover, the TDA-augmented RF models, more specifically the Pearson Landscape and wTO (Pearson Adjacency) Landscape models achieved perfect performance, with F1-scores and testing accuracies of 100%. However, other TDA representations, namely Distance Landscape, and wTO using Distance-based Adjacency underperformed compared to the other models, with an F1-score of 51% and 82% respectively. This suggesting that not all topological representations are equally informative for this classification task. These results demonstrate that while traditional ML models already perform exceptionally well, TDA-based models under optimal configurations can slightly surpass them.

Table 5.2 Performance Metrics of Raw and TDA Models for Cancer Type Prediction

	Model	F1-Score (%)	Training Accuracy (%)	Testing Accuracy (%)	MSE	Log Loss
Raw Data	LightGBM	98.00	100.00	97.60	0.10	0.13
	SVM	98.00	99.20	97.60	0.10	0.09
	NN	99.00	100.00	99.20	0.03	0.03
	RF	98.00	100.00	97.60	0.07	0.09
TDA-RF	Pearson Landscape	**100.00**	100.00	**100.00**	0.00	0.16
	Distance Landscape	51.00	100.00	58.40	1.64	2.13
	wTO (Pearson Adj) Landscape	**100.00**	100.00	**100.00**	0.00	0.06
	wTO (Distance Adj) Landscape	82.00	97.38	82.40	0.42	0.52

5.3.2 Cancer Staging Classification

Table 5.3 demonstrates the cancer staging results for the four stages of cancer of HNSC [2]. Unlike cancer typing prediction, we observe that the traditional machine learning models (SVM, NN, RF, LightGBM) underperformed significantly with F1-score ranging from 18-47% and testing accuracies between 35-49%. Despite the high training accuracies (95-97%), the F1-scores and testing accuracies are relatively low, suggesting signs of overfitting and a failure to generalize effectively to unseen data.

In contrast, TDA-based models demonstrate substantial improvements in performance. Most notably, wTO (Pearson Adjacency) achieved the highest test accuracy and F1-score of 79.80% and 78% respectively. This was also accompanied by relatively low error metrics (MSE = 1.61, Log Loss = 0.47). With wTO (Distance Adjacency) followed closely, with an F1-score of 75% and a testing accuracy of 74.75%, indicating reliable performance. However, not all TDA descriptors performed equally well as Pearson and Distance Landscapes achieved 51% and 42% F1-scores respectively.

Similar to the findings from cancer type classification, these results suggest that while topological representations can significantly enhance predictive performance, their effectiveness is highly dependent on the specific co-expression measures used to construct the simplicial complex.

[2]Classification results for cancer staging prediction can be found here: Cancer Staging Code and Results

Table 5.3 Performance Metrics of Raw and TDA Models for Cancer Staging Prediction

	Model	F1-Score (%)	Training Accuracy (%)	Testing Accuracy (%)	MSE	Log Loss
Raw Data	LightGBM	34.00	96.90	34.34	2.37	1.27
	SVM	18.00	95.87	35.35	1.10	1.26
	NN	37.00	97.93	37.37	2.33	2.05
	RF	47.00	95.87	49.50	1.68	1.18
TDA-RF	Pearson Landscape	51.00	100.00	51.52	2.27	0.89
	Distance Landscape	42.00	100.00	51.52	1.26	1.24
	wTO (Pearson Adj) Landscape	**78.00**	100.00	**79.80**	1.61	0.47
	wTO (Distance Adj) Landscape	**75.00**	99.48	**74.75**	1.68	0.58

5.3.3 Treatment Response Classification

Table 5.4 presents the treatment response prediction results for both raw and TDA-based models [3]. Among the traditional ML models, LightGBM achieved the highest F1-score (73%) with a testing accuracy of 75%, closely followed by RF (71% F1-score, 75% Testing Accuracy). SVM attained 66% F1-score and a testing accuracy of 76.73%, while NN reached an F1-score of 71% and a Testing Accuracy of 72.83%.

In contrast to the TDA-based models which consistently outperformed their raw-expression counterparts across all co-expression measures. Distance Landscape, Pearson Landscape and wTO (Pearson Adjacency) notably achieved exceptional performance with F1-scores of 100%, 98% and 93% respectively. Even the least successful TDA-based model wTO (Distance Adjacency) achieved an F1-score of 76%, surpassing all raw-expression models in this metric. These results highlight the robust predictive power of TDA-based descriptors.

Table 5.4 Performance Metrics of Raw and TDA Models for Treatment Response Prediction

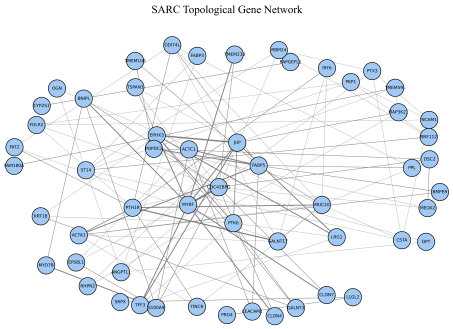
	Model	F1-Score (%)	Training Accuracy (%)	Testing Accuracy (%)	MSE	Log Loss
Raw	LightGBM	73.00	99.45	75.00	1.10	1.01
	SVM	66.00	76.09	76.73	0.82	0.51
	NN	71.00	97.51	72.83	1.11	0.68
	RF	71.00	97.78	75.00	0.97	0.47
TDA-RF	Pearson Landscape	**98.00**	100.00	**97.83**	0.98	0.20
	Distance Landscape	**100.00**	100.00	**100.00**	1.00	0.38
	wTO (Pearson Adj) Landscape	93.00	99.72	93.48	1.02	2.54
	wTO (Distance Adj) Landscape	76.00	100.00	75.00	1.35	0.90

[3] Classification results for treatment response prediction can be found here: Treatment Response Classification Code and Results

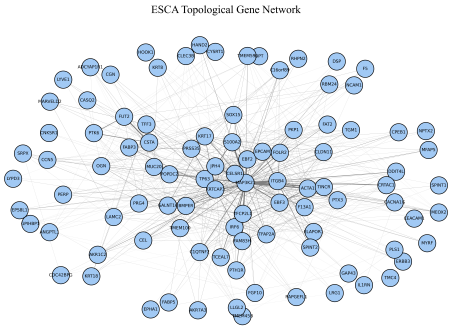
5.4 Biomarker Discovery

The results of the static visualization tool for each task and their corresponding phenotypes are presented below. These visualizations provide a clear depiction of topological gene interaction networks, highlighting key relationships and patterns relevant to biomarker discovery. Additionally, the interactive plots enhance interpretability by allowing users to interact directly with the network. The interactive visualizations can be accessed via the following link: https://nnyase.github.io/MSc-Thesis/. Moreover, the interactive visualization for each phenotype can be found in Supplementary Section A.1

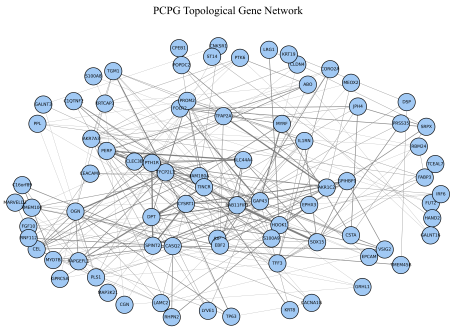
5.4.1 Cancer Types Biomarker Identification



(a) SARC



(b) ESCA



(c) PCPG

Fig. 5.4 Cancer Type Topological Networks.

5.4.1.1 Sarcoma

The sarcoma (SARC) network exhibited a compact yet robust structure, with hub genes such as *CDC42BPG*, *PTH1R*, *JUP*, *MUC20*, and *FABP5* dominating the interaction landscape (Figure 5.4a). Among these, *CDC42BPG* emerged as the most interactive gene, characterized by 17 topological connections, highlighting its central role in network dynamics. In this thesis, we will primarily focus

on *CDC42BPG* and *PTH1R* as key topological genes, given their important contributions to the structural integrity of the SARC network and their strong support in the existing sarcoma literature [156–160].

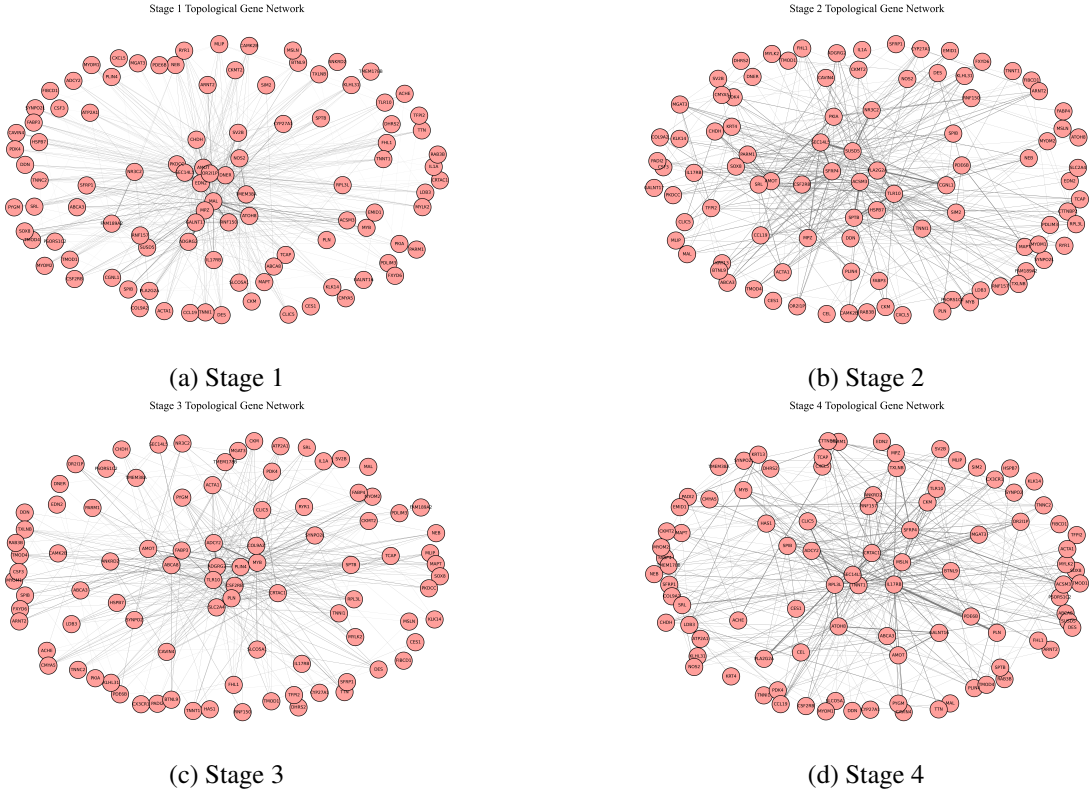
5.4.1.2 ESCA

The ESCA (Esophageal Cancer) topological gene network revealed a highly interconnected structure with several key hub genes playing central roles in WGTDA network organization (Figure 5.4b). Notably, genes such as *MAP3K2*, *ITGB4*, and *EpCAM* emerged as significant nodes with high centrality, indicating their critical roles in maintaining network stability and facilitating gene-gene interactions. In this thesis, we will focus specifically on two genes *MAP3K2* and *EpCAM* due to their substantial contributions to the network's organization and their well-documented importance in the ESCA literature [161–164].

5.4.1.3 PCPG

The PCPG (Pheochromocytoma and Paraganglioma) network (Figure 5.4c) appears slightly dense, reflecting a moderate level of interconnectivity among genes. Within this network, the key hub gene *RAB11-FTP* demonstrated notable connectivity, ranking second with 15 interactions. In this thesis, we will focus on *RAB11-FTP* due to its extensive connectivity and its relevance in PCPG-related literature [165].

5.4.2 Cancer Staging Biomarker Identification



(a) Stage 1

(b) Stage 2

(c) Stage 3

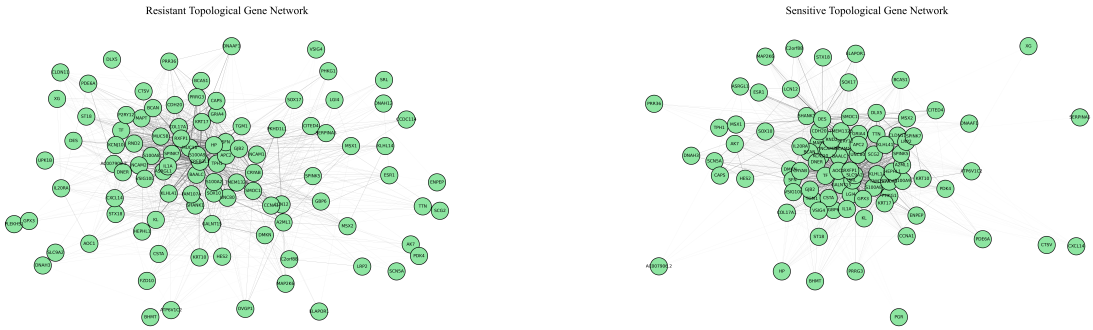
(d) Stage 4

Fig. 5.5 Cancer Staging Topological Network.

5.4.2.1 Stage IV HNSC

In our results of cancer staging for HNSC, we concentrated on Stage 4 due to its aggressive nature and the clear associations observed between topologically significant genes and their established roles in cancer biology. The Stage 4 network (Figure 5.5d) is relatively sparse, characterized by a limited number of highly connected nodes. Amid this sparse network, the *IL17RB* gene stands out as the most significant topological hub, with 33 connections identified across the network.

5.4.3 Treatment Response Biomarker Identification



(a) Resistance Network (b) Sensitive Network

Fig. 5.6 Treatment Response Topological Network

5.4.3.1 Resistant

The WGTDA network built from treatment-resistant or non-responsive samples is characterized by a dense web of highly interconnected genes (Figure 5.6a). Among the myriad interconnected nodes in this resistant network, *CREB3L1* emerged as the most prominent hub, exhibiting 16 topological interactions.

5.4.3.2 Sensitive

Similarly, the WGTDA network derived from treatment-sensitive or responsive samples (Figure 5.6b) also displays a high degree of connectivity, within this sensitive/responsive network, *HP* emerged as the central node, with 19 topological interactions.

Chapter 6

Discussion

This study set out to explore how persistent homology can help identify prognostic biomarkers and improve the prediction of clinical outcomes in gene expression data. Specifically, our discussion centers on the following objectives:

1. **Clinical Outcome Prediction:**

   (a) *Performance comparison between traditional ML models and TDA-based models:* We compare ML models trained on patient-specific topological landscapes to those trained on raw gene expression data, to assess any performance gain from the TDA-based models (Research Aim 1.2.4.1).

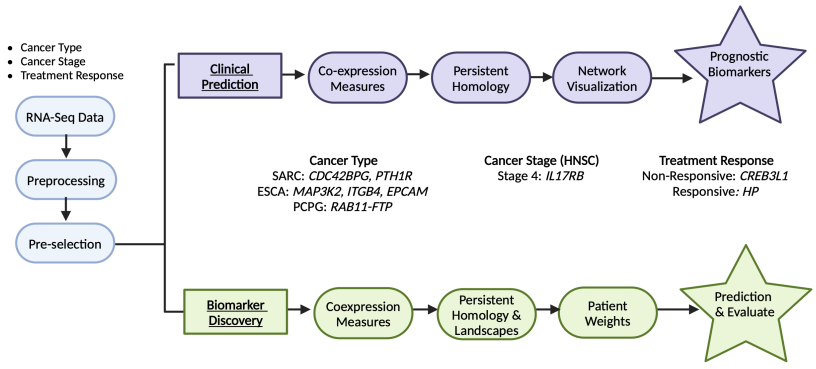
   (b) *Assess the impact of co-expression measures on subsequent model performance:* We investigate how different co-expression measures influence the construction of topological descriptors and their subsequent effect on model performance (Research Aim 1.2.4.2).

2. **Biomarker Discovery**

   (a) *Identification of Robust Biomarkers:* we evaluate the ability of WGTDA to detect robust biomarkers across various cancer types, stages, and treatment responses (Research Aim 1.2.3.1).

   (b) *Visualization and Interpretability:* We demonstrate the potential for a custom visualization tool to shed light on topological interactions, helping researchers interpret biomarker significance and gene–gene relationships (Research Aim 1.2.3.2).

By addressing these objectives, the discussion provides a comprehensive analysis of the potential of WGTDA for biomarker discovery and the integration of topological features for predictive modeling in clinical outcomes.



	Models	Cancer Type (%)	Cancer Stage (%)	Treatment Response (%)
TDA-RF	Pearson	100	51	98
	Distance	51	42	100
	wTO Pearson	100	78	93
	wTO Distance	82	75	76
Traditional ML	SVM	98	18	66
	NN	99	37	71
	LightGBM	98	34	73
	Random Forest	98	47	71

Table demonstrating the F1-scores for each clinical outcome prediction tasks, with the top two scores for each task highlighted.

Fig. 6.1 Overview of the proposed framework and key results. The workflow integrates RNA-Seq data preprocessing, feature selection, TDA for clinical outcome prediction and biomarker discovery. Notably, TDA-derived models outperformed traditional machine learning approaches across all clinical prediction tasks, with the highest F1-scores observed for cancer type, cancer stage, and treatment response. Additionally, biologically relevant biomarkers were identified using WGTDA in an agnostic manner and validated, demonstrating the framework's potential for uncovering prognostic molecular signatures.

6.1 Gene Set Pre-selection

The gene set pre-selection process, guided by *DESeq2* for identifying DEGs, played a critical role in refining the scope of downstream analyses. By isolating significantly up- and down-regulated genes (see Table 5.1), we focused on the most biologically relevant subset of the RNA-Seq data, a strategy which helped reduce noise and computational overhead. By filtering for the most relevant DEGs, subsequent TDA-based modeling benefits from a richer yet tractable feature space, potentially improving the specificity of downstream classification and biomarker identification.

6.2 Exploratory Data Analysis

6.2.1 Cancer Type EDA

The PCA, t-SNE and UMAP plots in Section 5.2 demonstrate a clear separability, providing strong evidence for the diversity in transcriptomic profiles across the three cancer types. This observation is particularly important for downstream machine learning tasks, as the presence of distinct clusters suggests that the classification models will be able to learn meaningful boundaries between cancer types. Consequently, we hypothesize that predicting cancer types will be the least challenging classification problem within this thesis.

6.2.2 Cancer Staging EDA

The staging results from PCA, t-SNE, and UMAP collectively indicate considerable heterogeneity and poor class separability between stages I-IV. All three dimensionality reduction techniques highlight substantial overlap between stages, suggesting that differences in gene expression profiles across stages are subtle. This observed overlap has critical implications for downstream machine learning tasks. The lack of distinct separation between phenotypic stages implies that both TDA-enhanced and traditional classification models will face challenges in learning separable decision boundaries.

6.2.3 Treatment Response EDA

For treatment response, results across PCA, t-SNE, and UMAP emphasize the difficulty of separating resistant and sensitive phenotypes, stemming from overlapping gene expression profiles and class imbalance. From a machine learning perspective, the class imbalance introduces an additional challenge, as models may inherently favor the majority class (sensitive) over the minority class (resistant) [166]. Moreover, the limited separability observed through EDA suggests that both traditional models and TDA-based models may struggle to differentiate between responders and non-responders.

In summary, the EDA findings suggest that Task 1 (Cancer Type) will likely be the least challenging, given the clear separations observed among the different cancer types. By contrast,

Task 2 (Cancer Staging) and Task 3 (Treatment Response) present significantly more overlap in their underlying data, making them inherently more difficult. The added complexity of class imbalance in Task 3 further amplifies these challenges, as ML models may struggle to capture the minority class effectively.

6.3 Clinical Outcome Prediction

In this section, we applied persistent homology to predict clinical outcomes, including cancer type, staging, and treatment response. By leveraging topological features from gene co-expression networks, we demonstrated that TDA improves classification performance in complex tasks, particularly in modeling tumor stage progression and treatment response.

6.3.1 Cancer Type Classification

For cancer type prediction, data reduction methods, such as PCA, t-SNE and UMAP revealed that the cancer types were relatively well-separated in lower-dimensional spaces, providing an early indication that the classification would be relatively straightforward. This insight from the EDA was confirmed by the high performance results for cancer type prediction (Table 5.2), where both traditional ML models (SVM, NN, LightGBM and RF) and TDA-based models achieved excellent test accuracies and F1-scores. Given this clear separability and the strong performance of traditional ML models, one could argue that the use of TDA or any other data representation technique might not have been strictly necessary for this specific task.

However, the inclusion of TDA serves a clear purpose. It establishes an important baseline comparison between conventional ML models and TDA-enhanced models, even in datasets where gene expression profiles are easily separable.

The Pearson and wTO (Pearson Adjacency) models achieved perfect results with 100% F1-scores and test accuracy, surpassing traditional ML models by a slight but notable 1-2% margin. This result demonstrates that even in tasks where traditional models excel, TDA can still provide slightly better margins.

This marginal improvement, while numerically small, can be viewed from two perspectives. On one hand, some may argue that even a 1–2% difference in F1-score is significant, particularly in the context of clinical applications where a small performance gain could translate into more accurate diagnoses and improved patient outcomes. On the other hand, this improvement can also be seen as trivial and negligible given the already excellent performance of traditional models and the relative straightforward nature of this classification task.

Nevertheless, the results from Distance and wTO (Distance Adjacency) RF introduce an important caveat. These models underperformed significantly compared to both traditional models and their TDA counterparts based on Pearson Correlation. This suggests that the choice of co-expression measure has a profound impact on the reliability and strength of the model's performance. Distance Correlation may not be well-suited for generating TDA descriptors for gene expression data due to its inability to provide directional information of the gene-gene relationship. Furthermore, while Distance Correlation can account for non-linear dependencies, it may inadvertently introduce spurious correlations which may find signals that are not biologically related. These false signals can introduce noise into the topological features, ultimately weakening the performance of the TDA-based models. Given these findings, Pearson Correlation and wTO with Pearson-based adjacency both yielded superior results for cancer type prediction, indicating their effectiveness in constructing topological descriptors that lead to stronger downstream model performance.

As such, while these results highlight TDA's potential to refine predictive performance, the true test of its utility will lie in the subsequent tasks of cancer staging and treatment response prediction. These tasks are inherently more complex, with less separable classes and more subtle patterns to detect. Success in these tasks would provide stronger evidence of TDA's capacity to offer meaningful improvements over conventional approaches, justifying its computational and interpretative overhead.

6.3.2 Cancer Staging Classification

Unlike the clear separation observed in cancer types, EDA for cancer staging suggested that staging would present a more challenging problem as there was a poor separability observed for the different stages. This challenge became evident in the performance of non-TDA models which underperformed significantly with F1-scores ranging from 18–47% (Table 5.3). Although these models had high training accuracies (95–97%), they failed to generalize well on unseen data, indicating substantial overfitting. This suggests that traditional ML models struggled to capture the subtle patterns and complex relationships underlying tumor progression.

In contrast, both wTO (Pearson and Distance Adjacency) significantly outperformed traditional ML, achieving an F1-score of 78% and 75% respectively. This is an improvement of 25-40% over traditional methods. These results highlight the predictive strength of topological features in capturing meaningful patterns that raw gene expression data fail to reveal, particularly in cases where traditional models overfit.

However, similar to the cancer type task not all co-expression measures were equally effective as Pearson and Distance Landscapes achieved 51% and 42% for F1-scores respectively. wTO-based methods performed significantly better, highlighting the importance of selecting the appropriate co-expression measure. The superior performance of wTO suggests that shared neighborhood

information among genes offers a more nuanced representation to distinguish complex phenotypes like cancer staging. This finding aligns with Mashatola et al. [22], who demonstrated that topological signatures derived from signed-TOM improved cancer phenotype prediction accuracy by nearly 20% compared to the commonly used Distance Correlation metric. These results emphasize that wTO-based TDA models outperform both traditional ML models and other co-expression approaches, reinforcing the value of neighborhood-aware co-expression in capturing dependencies associated with complex phenotypes.

Given the inherent complexity of tumor progression, topological features offer a structured way to capture subtle gene expression changes that ML models can leverage more effectively than the raw expression data itself. Testing this advantage becomes even more crucial in treatment response, where accurately predicting whether a patient will respond to treatment carries significant clinical value.

6.3.3 Treatment Response Classification

EDA for treatment response indicated substantial overlap in gene expression profiles between the sensitive and resistant groups, foreshadowing the challenges of achieving clear separation in downstream prediction tasks. This complexity was further compounded by class imbalance in which there were far more sensitive samples than resistant samples, making it difficult for classifiers to accurately capture the minority class.

Despite these obstacles, TDA-based descriptors consistently outperformed the raw-expression models. Even the least successful TDA model, wTO (Distance Adjacency), which achieved a 76% F1-score which surpassed all the raw-expression models, more notably the best-performing traditional ML model LightGBM (73% F1-score). This reinforces the advantage of incorporating topological features in capturing meaningful patterns that conventional methods struggle to detect (Table 5.4).

At the higher end of performance, the Distance Landscape delivered perfect performance with an F1 score of 100% and a test accuracy of 100%. Following closely is Pearson and wTO (Pearson Adjacency) with an F1 score of 98% and 93% respectively. These results further highlight the effectiveness of TDA-based approaches in clinical outcome prediction, demonstrating their ability to improve traditional ML models in gene expression tasks.

A key concern in the treatment response data was the class imbalance between sensitive and resistant samples. In such scenarios, relying solely on accuracy can be misleading, as models may bias predictions towards the majority class to inflate their apparent performance, without meaningfully capturing minority class instances. To address this concern, F1-score was used as a primary evaluation metric, emphasizing a balance of precision and recall. This approach is especially important when

minority class predictions (e.g., treatment resistance) carry significant clinical implications.

Nonetheless, the performance from the TDA-driven models suggest that topological fingerprints improve ML model performance even in minority class instances. This insight resonates with broader realities in biological and clinical research, where datasets are often imbalanced, and sample sizes are usually limited.

6.3.4 Summary of Clinical Outcome Prediction

The results from all clinical prediction tasks—cancer type, cancer staging, and treatment response—demonstrate the strong predictive advantage of TDA-based models over traditional ML approaches, particularly in more complex classification problems.

In cancer type classification (Table 5.2), the traditional ML models performed well, aligning with the clear separation observed in the EDA. Although TDA-based models slightly improved performance (1-2%), their use was not strictly necessary for this task due to the high performance of the traditional ML models. Moreover, Distance Landscapes did not perform well suggesting that the choice of co-expression measure is imperative.

For cancer staging classification (Table 5.3), traditional models struggled, with F1-scores ranging between 18% and 47%, indicating severe overfitting and poor generalization. In contrast, wTO-based TDA models significantly outperformed traditional approaches, achieving up to 78% F1-score. This highlights the importance of shared gene-gene neighborhood information, as these topological descriptors provided features that are more informative representation for ML classification.

Moreover, treatment response classification (Table 5.4) was particularly challenging due to substantial class imbalance and overlapping in gene expression patterns, as observed in the EDA. Despite this, TDA-based models consistently outperformed traditional ML approaches, with even the lowest performing TDA model, wTO (Distance Adjacency) surpassing the all the traditional models. These results demonstrate that topological fingerprints capture richer structural information, allowing ML models to detect meaningful patterns that differentiate complex phenotypes, even in highly imbalanced datasets.

Based on these findings, we recommend adopting Pearson Correlation or wTO (Pearson-based adjacency) as standard approaches for constructing topological descriptors from gene expression. We recommend Pearson for its widespread use and popularity as a correlation measure. However, for more nuanced phenotypes, such as tumor staging or treatment response, we further recommend wTO with

Pearson-based adjacencies, which has shown superior performance in these complex heterogeneous phenotypes.

6.4 Biomarker Discovery

In this section, we applied WGTDA to identify potential biomarkers across multiple phenotypes, including cancer types, cancer staging, and treatment response. By focusing on topological features (β numbers) in gene–gene interaction networks, WGTDA highlighted key genes without relying on pre-existing biological assumptions. Despite this agnostic approach, the identified biomarkers were strongly corroborated by existing literature.

6.4.1 Cancer Type Biomarker Identification

6.4.1.1 Sarcoma (SARC)

The WGTDA network for SARC (Figure 5.4a) highlights *CDC42BPG* and *PTH1R* as key topological genes, identified as the most central nodes based on their persistence and connectivity within the network. Both genes are proposed biomarkers for sarcoma in literature [156–160]

The significance of *CDC42BPG* as a top-ranking biomarker is strongly supported by the experimental findings of Jayabal et al. [156]. Their study investigated the NELL2–Robo3 signaling pathway in Ewing sarcoma cells and demonstrated its critical role in regulating *CDC42* activity. Using RNA interference and pull-down assays they knocked down or silenced *NELL2* or its receptor Robo3, which led to a significant increase in *CDC42* activity. As a result, *CDC42* disrupted the assembly and stability of BAF (*BRG1/BRM*-associated factor) complexes, which are important for controlling gene expression. However, when *CDC42* was blocked with an inhibitor, BAF complexes stabilized, and normal cell growth returned. These findings clearly demonstrate that *CDC42* activity promotes Ewing sarcoma cell growth by altering chromatin-remodeling processes.

Likewise, *PTH1R*, also identified through WGTDA plays a critical role in osteosarcoma (OS) progression, as the review by Al et al. [157] reported *PTH1R* overexpression correlates with poor prognosis, promoting increased metastasis, chemoresistance, tumor growth, and reduced survival. Experimental evidence supporting these findings includes immunohistochemical (IHC) studies in canine OS tissues, which revealed that dogs with tumors showing strong *PTH1R* staining had significantly shorter survival times—approximately half as long (median 212 days) as those with weak staining (median 459 days) [158]. Furthermore, recent experiments by Liaoning Cancer Hospital demonstrated that treating human Saos-2 and U2OS OS cell lines with mangiferin, resulted in a significant decrease in both *PTH1R* mRNA and protein expression, as confirmed by immunofluorescence. This downregulation was accompanied by inhibited proliferation, migration, and invasion of OS cells, thereby suggesting that suppression of *PTH1R* is directly correlated with reduced tumor aggressiveness

and supporting its potential as a prognostic biomarker in sarcoma [159]. In a similar study by Li et al. (2019) [160], quercetin-induced inhibition of *PTH1R* on proliferation, migration, and invasion in U2OS and Saos-2 cells. These studies reinforce *PTH1R* value as a prognostic biomarker and potential therapeutic target in sarcoma.

Although WGTDA has successfully identified *PTH1R* and *CDC42* as key biomarkers and therapeutic targets in sarcoma through topological analysis, its true potential lies in broader validation. Demonstrating its ability to uncover biomarkers across additional tumor types, such as ESCA and PCPG, as well as in other clinically relevant phenotypes, including cancer staging and treatment response, will ultimately confirm WGTDA as a powerful tool for biomarker discovery in precision oncology.

6.4.1.2 Esophageal Carcinoma (ESCA)

For ESCA, WGTDA identified *MAP3K2* and *EpCAM* as topologically significant genes (Figure 5.4b). The following literature corroborates their roles as ESCA biomarkers, reinforcing WGTDA's capability to uncover clinically relevant markers and highlights its potential for phenotype-driven discovery in oncology.

MAP3K2 plays a key role in the MAPK signaling pathway, which regulates cell proliferation and survival. Notably through IHC analysis of *MAP3K3*, a closely related kinase, in 93 Esophageal squamous cell carcinoma (ESCC) samples revealed significant overexpression in both dysplasia and tumor tissues compared to normal mucosa [164]. Higher *MAP3K3* levels correlated with reduced disease-free survival (median 10 vs. 19 months, $p = 0.04$). These findings suggest that *MAP3K2*, like *MAP3K3*, may drive ESCC progression and serve as a prognostic biomarker for ESCA.

Similarly, *EpCAM* has been extensively studied in the context of ESCA. Firstly, Kimura et al. (2007) [161] analyzed 138 esophageal cancer samples using RT-PCR, IHC, and ELISA (Enzyme-Linked Immunosorbent Assay), confirming that *EpCAM* expression was significantly higher in tumor tissues compared to normal tissues ($p < 0.0001$). Moreover, *EpCAM* expression was correlated with tumor depth, stage, blood-vessel invasion and infiltrative growth pattern for ESCA. Moreover, Stoecklein et al. (2006) [163] further reinforced *EpCAM's* role in ESCA by analyzing 70 primary ESCC samples using IHC. Their findings revealed *EpCAM* neo-expression in 79% of tumors, with higher expression levels (3+) significantly correlating with decreased relapse-free survival ($p = 0.0001$) and overall survival ($p = 0.0003$). Further evidence from Matsuda et al. (2014) [162] demonstrated through IHC in 74 ESCC patients that *EpCAM* overexpression correlates with poor survival ($p = 0.026$) and is an independent prognostic factor ($p = 0.004$). Functional experiments using TE4, TE10, and TE14 ESCC cell lines confirmed that *EpCAM* knockdown suppressed proliferation by downregulating *CCND1* and *CCNE2*, key regulators of the cell cycle. Additionally, *in-vivo* studies showed that

EpCAM knockdown reduced tumorigenesis, with tumors forming in only 2 out of 10 mice, compared to 7 out of 10 in controls, highlighting *EpCAM's* role as both a therapeutic target and prognostic biomarker in ESCC.

With WGTDA successfully identifying key biomarkers in both SARC and ESCA, its ability to reveal clinically significant genes is reinforced. The next step will involve extending this analysis to PCPG and more complex phenotypes such as cancer staging and treatment response, further demonstrating WGTDA's potential for phenotype-driven discovery in precision oncology.

6.4.1.3 Pheochromocytoma and Paraganglioma (PCPG)

In PCPG, RAB11-FIP1 emerged as a key node with 15 network connections (Figure 5.4c), aligning with literature that highlights the role of RAB pathways in tumorigenesis. Rab11, a small GTPase involved in vesicle trafficking, has been linked to metastatic PCPG through mutations in MYO5B, a motor protein that interacts with Rab11, disrupting vesicle transport and promoting tumor progression. Additionally, altered Rab11 and RAB11-FIP1 expression distinguishes PCPG subtypes, reinforcing their potential as prognostic biomarkers [165].

Taken together, our findings highlight the power and potential of WGTDA. Despite making no prior biological assumptions, WGTDA consistently pinpointed genes (e.g., *CDC42BPG*, *PTH1R*, *MAP3K2*, *EpCAM*, *RAB11-FIP1*) that are strongly supported by experimental and clinical evidence across multiple cancer types. This not only validates WGTDA's capability in uncovering meaningful biomarkers but also lays the groundwork for its broader application in diverse phenotypic analyses (cancer stage and treatment response), which will be further explored.

6.4.2 Cancer Staging Biomarker Identification

6.4.2.1 Stage IV HNSC

For stage IV HNSC (Figure 5.5d), WGDTA pinpointed *IL17RB* as the most significant topological biomarker, aligning with recent findings by Sun et al. (2023) [167]. Their study analyzed TCGA HNSC cohort, revealing that HPV-positive tumors exhibited higher *IL17RB* expression, which correlated with increased viral load and an improved prognosis. The researchers found that *IL17RB* expression was associated with an elevated presence of memory B cells and activated NK cells, two immune populations linked to enhanced anti-tumor immunity. Kaplan-Meier survival analysis confirmed that patients with high *IL17RB* expression had significantly better clinical outcomes compared to those with lower expression.

6.4.3 Treatment Response Biomarker Identification

6.4.3.1 Resistant

Our own WGTDA approach independently flagged *CREB3L1* as a pivotal node in the resistant or non-responsive network (Figure 5.6a). Notably, this computational discovery aligns with the findings of Denard et al. (2018), who established a mechanistic and clinical link between *CREB3L1* expression and sensitivity to doxorubicin-based chemotherapy in Triple-Negative Breast Cancer (TNBC) [168]. Denard et al. demonstrated that high CREB3L1 expression (IRS ≥ 4) correlates significantly with positive therapeutic responses to doxorubicin-based regimens, as quantified by the Residual Cancer Burden (RCB) system. Conversely, tumors exhibiting low CREB3L1 expression (IRS ≤ 3) were predominantly resistant to treatment, aligning with higher RCB class scores indicative of poor clinical outcomes. The identification of *CREB3L1* through WGTDA and its subsequent alignment with established clinical evidence provides a compelling example of the methodology's power and reliability.

6.5 Limitations of Present Study

In this section, we critically evaluate the inherent constraints of our research, particularly those related to clinical outcome prediction and biomarker discovery using topological features. While the methods employed offer promising avenues for future exploration, the following limitations highlight areas where caution is warranted, and improvements are needed.

6.5.1 Clinical Outcome Prediction

1. *Interpretability of Topological Features:*

   While persistent landscapes provide a way to integrate topological information in machine learning models, they lack intuitive interpretability [51]. Once the original topological signatures are transformed into the persistent landscape representation, it becomes exceedingly difficult to map them back to the underlying simplices or genes. Although one can still perform variable importance analyses on these derived features, the aggregated nature of persistent landscapes limits the ability to pinpoint precisely which genes drive a given prediction. This loss of direct mapping compromises our ability to gain biologically meaningful insights and hinders the development of clinically actionable hypotheses.

6.5.2 Biomarker Discovery

1. *Limited Gene Search Space:*

   A key limitation in our biomarker discovery pipeline is the reliance on differential gene expression analysis to reduce the dataset to a smaller set of significantly altered genes. While

this strategy mitigates the computational burden associated with computing the persistent homology for WGTDA, it inherently excludes the vast majority of the 19,000 protein-coding genes. This exclusion may lead to overlooking potentially critical genes or pathways that do not exhibit strong differential expression yet still play a pivotal role in disease etiology or treatment response. By narrowing our focus prematurely, we risk missing novel biomarkers that could only emerge when analyzing the entire genomic landscape. Although this compromise ensured computational feasibility, it restricts our findings and may limit their translational potential. Future studies could address this by leveraging more scalable computational approaches or by integrating a broader range of genes in staged analyses. Future studies could address this limitation by leveraging emerging computational paradigms such as quantum computing. More specifically, quantum TDA techniques have demonstrated linear-depth computation with exponential scaling advantages theoretically allowing us to encode more features than we can on a classical computer [49, 169].

2. *Treatment Response Dataset*

   Another significant shortcoming stems from the decision to combine multiple anatomical regions and cancer types into a single treatment response dataset. Although pooling various phenotypes initially appeared to increase the size of the dataset, and widen the scope of our analysis, it also introduced substantial heterogeneity. This heterogeneity obscures condition-specific signals and makes it more challenging to draw definitive conclusions about any cancer type or treatment response. In hindsight, a more focused approach such as restricting the analysis to a single phenotype's response to treatment or evaluating the effects of one specific therapeutic intervention could have produced clearer insights and more robust biomarkers. Such an approach would have minimized the confounding effects introduced by inter-cancer variability.

6.6 Future Work

1. *Toward an Integrated WGTDA Framework and Visualization Platform:*

   A natural progression of this research involves transforming the existing network visualization tool into a comprehensive application for both researchers and bioinformaticians. This platform would enable users to:

   (a) **Upload and Preprocess Data:** Provide a dashboard where researchers can easily upload gene expression data or differential expression results.

   (b) **Seamless WGTDA Computations:** Run the full WGTDA pipeline behind the scenes, shielding users from the complexities of persistent homology and code. This will facilitate the generation of topological features—such as β_1 and β_2 without requiring advanced mathematical or computer science knowledge.

(c) **Interactive Network Exploration:** Offer a dynamic, web-based interface where users can click on nodes or edges to view gene annotations and interaction strengths. Incorporating graph-based algorithms (e.g., clustering methods) would help identify key sub-networks and significant modules within the data.

(d) **On-the-Fly Functional Enrichment:** Integrate real-time Gene Set Enrichment Analysis (GSEA) to reveal how the topologically significant genes are distributed across various functional terms. This enrichment step could be complemented by Gene Ontology (GO) [170], Kyoto encyclopedia of genes and genomes (KEGG) [171] , and Reactome [172] annotations to dynamically color, size, or shape network nodes based on functional categories.

(e) **Customizable Visualizations:** Provide options to adjust layout parameters and highlight nodes or edges based on user-defined criteria. Genes involved in particular pathways such as immune response could be easily distinguished by color or other visual attributes.

By integrating these features into a unified network biology interface, researchers would gain a powerful, user-friendly tool. This enhanced platform would seamlessly bridge the gap between raw high-throughput data, TDA, and functional interpretation, ultimately driving deeper insights into complex biological systems.

2. *Towards Quantum Topological Data Analysis (QTDA):*

A promising avenue to overcome the prohibitive computational costs of large-scale TDA lies in leveraging emerging quantum computing technologies. TDA requires systematically exploring high-dimensional data structures—a task that expands exponentially with each additional simplices (or gene). Even high-performance GPUs can become overwhelmed when datasets expand beyond a few hundred simplices.

In contrast, quantum computing harnesses phenomena such as superposition and entanglement to process multiple states in parallel, potentially offering exponential speedups for TDA algorithms. As quantum devices, such as IBM's "Heron" series, steadily increase their qubit counts and refine circuit depth [173–175], the feasibility of computing persistent homology for thousands of genes and computing higher-order β numbers (*betti-3,4,5*) becomes more than just a theoretical possibility. Meanwhile, recent theoretical advances [48, 49, 169] have introduced quantum TDA frameworks that optimize both algorithmic steps and circuit complexity, further enhancing scalability.

These developments pave the way for "quantum WGTDA" (QWGTDA), transforming computations once deemed infeasible into tasks that could be completed within hours—or even minutes. By providing the ability to probe the full genomic space rather than restricting analyses to a few

hundred genes and computing β_3 and higher, QWGTDA opens unprecedented opportunities for identifying novel biomarkers and elucidating the complex mechanisms underlying disease.

3. *Interpretation and Validation through In-vivo and In-vitro Experiments:*

   To further enhance the robustness and translational relevance of our TDA-derived findings, we recommend establishing a closer collaboration with biologists and clinical experts. Such interdisciplinary engagement is essential to ensure that the topological features are further validated and accurately interpreted within the biological context.

   Future research should incorporate targeted *in-vitro* and *in-vivo* experiments to investigate the biological functions and interactions of the gene sets identified through WGTDA. By aligning these experimental validations with known biological processes, researchers can deepen the understanding of how these genes contribute to diseases, ultimately reinforcing their potential as novel prognostic and therapeutic targets.

   This integrated approach, which bridges advanced computational methods with experimental biological validation, will not only bolster the credibility of TDA-derived biomarkers but unravel the potential of TDA in advancing precision oncology.

4. *Integrating Harmonic Representative Homology:*

   In WGTDA, topological cycles are identified through boundary matrix reduction, yet this approach fails to distinguish which simplices within the "hole" are most essential. Harmonic persistent homology offers a potential solution by assigning weights or “harmonic coefficients” to individual simplices, thereby quantifying their contribution to the homology group [145, 176]. In other words, rather than viewing each cycle as an unweighted set of simplices, harmonic homology highlights those simplices whose removal would destroy the cycle’s topological feature altogether. This refinement could provide deeper insight into the underlying biological structures, since simplices with higher harmonic coefficients would likely represent critical gene within a homology class. By incorporating harmonic representative homology into WGTDA, we could move beyond simply identifying the presence of a cycle and begin to characterize its most influential elements—an advancement that holds promise for biomarker discovery, where the identification of pivotal interactions may lead to better biomarkers.

Chapter 7

Conclusion

We began this thesis with the fundamental observation that biological data has shape, and that this shape matters. Beneath the high-dimensional and often noisy landscape of gene expression profiles lies hidden topological structures (loops, holes, and voids) that capture the intricate relationships driving cancer development and progression. By embracing this perspective, we placed TDA and persistent homology at the core of a novel analytical framework designed to address two intertwined challenges in cancer research: clinical outcome prediction and biomarker discovery.

7.1 Clinical Outcome Prediction

Our findings strongly support the integration of TDA into clinical outcome prediction, demonstrating its value in capturing nuanced patterns that allow ML methods to learn more effectively. In challenging tasks like cancer staging and treatment response, where conventional ML models often overfit and struggle to generalize due to class imbalance, TDA-based models have consistently demonstrated superior performance.

For example, in cancer staging, although traditional ML models achieved high training accuracies, they failed to generalize on unseen data, demonstrating substantial overfitting. In contrast, wTO-based TDA models consistently outperformed these traditional approaches, delivering a 25–40% improvement in F1 performance, highlighting the utility of TDA-based models.

Notably, the topological fingerprints demonstrated robustness in handling class imbalance and overlapping gene expression patterns. This was evident as the lowest-performing TDA model (wTO using Distance-based adjacency), surpassed all traditional ML models in treatment response. This further demonstrates the clinical relevance of these methods, particularly in predicting critical minority classes like treatment resistance.

Moreover, the choice of co-expression measures emerged as a pivotal factor in subsequent model performance. Distance Correlation, which lacks directional information, performed poorly in certain tasks, highlighting its limitations in capturing meaningful interactions that can be exploited by ML models. In contrast, shared neighborhood measures such as wTO using Pearson-based adjacency consistently provided better model performance. These results emphasize the importance of shared gene-gene neighborhood information in generating a meaningful representation of complex phenomena like tumor progression and treatment response.

Based on these results, we recommend using Pearson Correlation as the standard approach for constructing topological descriptors, given its widespread use and proven effectiveness, and wTO (Pearson-based adjacency) for more nuanced and complex phenotypes like tumor staging and treatment response.

7.2 Biomarker Discovery

In this study, we applied WGTDA to identify prognostic biomarkers across various cancer phenotypes, including cancer type, stage, and treatment response. We further introduced a visualization framework that highlights gene–gene interactions derived from topological features, providing biological insight and interpretability for identifying biomarkers. The results demonstrate WGTDA's clinical potential for biomarker discovery, as it uncovers robust candidate genes without relying on pre-existing biological assumptions. Despite its agnostic approach, the biomarkers uncovered by WGTDA were consistently validated against experimental and clinical evidence, underscoring the method's reliability and potential clinical utility.

For cancer type, WGTDA pinpointed *CDC42BPG* and *PTH1R* as key biomarkers in sarcoma, with the former being implicated in Ewing sarcoma cell growth [156] and the latter in osteosarcoma (OS) progression [157, 160, 158, 159]. Moreover, *MAP3K2* and *EpCAM* were identified as significant biomarkers for esophageal carcinoma (ESCA) [164, 163, 162], whereas *RAB11-FTP* was implicated in Pheochromocytoma and Paraganglioma (PCPG) [165]. In cancer staging, *IL17RB* was identified as a significant biomarker for stage IV head and neck squamous cell carcinoma (HNSC), with higher expression levels correlating with improved prognosis [167]. Lastly, for treatment response, *CREB3L1* emerged as a pivotal biomarker in doxorubicin-resistant triple-negative breast cancer (TNBC), with its expression levels directly linked to chemotherapy sensitivity and patient outcomes [168].

The results confirm WGTDA's capability in identifying biologically and clinically meaningful biomarkers across different cancer phenotypes. This reinforces its potential as a relevant framework for biomarker discovery in precision oncology. Furthermore, the ability of WGTDA to extract topological features without prior biological assumptions provides a significant advantage over traditional biomarker discovery methods, allowing for unbiased identification of novel prognostic and

therapeutic targets.

This thesis has demonstrated the power of TDA in advancing biomarker discovery and enhancing clinical outcome predictions in cancer research. Beyond these contributions, the work opens promising avenues for future research, including the development of a comprehensive WGTDA toolkit and platform, the integration of quantum TDA algorithms to increase scalability, further validation through *in-vivo* and *in-vitro* experiments, and the incorporation of harmonic representative homology to identify essential simplices in topological features.

7.3 Troubleshooting

During this project, several challenges emerged that provided valuable insights into the practical implementation of TDA methods. One notable challenge was developing a custom TDA package in C++ that included implementations for both the Vietoris–Rips complex and persistent homology. To facilitate broader accessibility and integration with existing Python-based workflows, Pybind11 was employed to port the C++ code to Python. While the Vietoris–Rips component appeared to be correct, the computed persistent homology results differed significantly from those produced by established packages like Gudhi and maTILDA. Despite thorough investigation, the discrepancies remained unresolved, suggesting that the persistent homology computation did not function as intended. Although this effort ultimately fell short of its intended goals, it offered a valuable opportunity to revisit C++ and strengthen proficiency in the language.

7.4 Code Availability

All code developed and used for this project is openly accessible on GitHub. The primary repository, which contains the data processing scripts, models, results for clinical outcomes, network visualizations, and other supporting materials for the MSc thesis, can be found at:

- https://github.com/nnyase/MSc-Thesis

Additionally, the Weighted Gene Topological Data Analysis (WGTDA) tool the topology-based framework for gene expression data analysis was open-sourced by IBM Research Africa and is available at:

- https://github.com/IBM/WGTDA

Both repositories include detailed documentation, sample workflows, and instructions for setup and usage. Any updates or improvements to the code will be committed to these public repositories to ensure ongoing accessibility and reproducibility.

7.5 Data Availability

This thesis leverages publicly available datasets from The Cancer Genome Atlas (TCGA) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC). Specifically, gene expression data for sarcomas (SARC), esophageal carcinomas (ESCA), paragangliomas/pheochromocytomas (PCPG), and head and neck squamous cell carcinoma (HNSC) were obtained through TCGA (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). Proteomic and genomic data for treatment response analysis were sourced from CPTAC (https://proteomics.cancer.gov/programs/cptac). Further details and references for each dataset can be found within the text.

Bibliography

[1] George Stiny. *Shape: talking about seeing and doing*. MIt Press, 2006.

[2] Pek Y Lum, Gurjeet Singh, Alan Lehman, Tigran Ishkanov, Mikael Vejdemo-Johansson, Muthu Alagappan, John Carlsson, and Gunnar Carlsson. Extracting insights from the shape of complex data using topology. *Scientific reports*, 3(1):1236, 2013.

[3] Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5(1):501–532, 2018.

[4] Amrita Chattopadhyay and Tzu-Pin Lu. Gene-gene interaction: the curse of dimensionality. *Annals of translational medicine*, 7(24), 2019.

[5] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of big data*, 2:1–21, 2015.

[6] Zulema Udaondo. Big data and computational advancements for next generation of microbial biotechnology. *Microbial Biotechnology*, 15(1):107, 2022.

[7] A. Garin and G. Tauzin. A topological "reading" lesson: classification of mnist using tda. *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2019.

[8] Pablo G Cámara. Topological methods for genomics: present and future directions. *Current opinion in systems biology*, 1:95–101, 2017.

[9] Rebecca L Siegel, Angela N Giaquinto, and Ahmedin Jemal. Cancer statistics, 2024. *CA: a cancer journal for clinicians*, 74(1), 2024.

[10] Euna Jeong and Sukjoon Yoon. Current advances in comprehensive omics data mining for oncology and cancer research. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, page 189030, 2023.

[11] Nivedhitha Mahendran, PM Durai Raj Vincent, Kathiravan Srinivasan, and Chuan-Yu Chang. Machine learning based computational gene selection models: a survey, performance evaluation, open issues, and future research directions. *Frontiers in genetics*, 11:603808, 2020.

[12] Robert Clarke, Habtom W Ressom, Antai Wang, Jianhua Xuan, Minetta C Liu, Edmund A Gehan, and Yue Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature reviews cancer*, 8(1):37–49, 2008.

[13] David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnér. Data integration in the era of omics: current and future challenges. *BMC systems biology*, 8:1–10, 2014.

[14] Ndivhuwo Nyase, Lebohang Mashatola, Aviwe Kohlakala, Kahn Rhrissorrakrai, and Stephanie Muller. Wgtda: A topological perspective to biomarker discovery in gene expression data. *arXiv preprint arXiv:2402.08807*, 2024.

[15] Ndivhuwo Nyase and Lebohang Happy Mashatola. WGTDA. 2024.

[16] Attila A Seyhan. Biomarkers in drug discovery and development. *Eur Biopharm Rev*, 5:19–25, 2010.

[17] Theodosia Charitou, Kenneth Bryan, and David J Lynn. Using biological networks to integrate, visualize and analyze genomics data. *Genetics Selection Evolution*, 48:1–12, 2016.

[18] Daniele Merico, David Gfeller, and Gary D Bader. How to visually interpret biological data using networks. *Nature biotechnology*, 27(10):921–924, 2009.

[19] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[20] Andreas Krämer, Jeff Green, Jack Pollard Jr, and Stuart Tugendreich. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4):523–530, 2014.

[21] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

[22] Lebohang Mashatola, Zubayr Kader, Naaziyah Abdulla, and Mandeep Kaur. Enhancing the vietoris–rips simplicial complex for topological data analysis: applications in cancer gene expression datasets. *International Journal of Data Science and Analytics*, pages 1–18, 2024.

[23] Anuraag Bukkuri, Noemi Andor, and Isabel K Darcy. Applications of topological data analysis in oncology. *Frontiers in artificial intelligence*, 4:659037, 2021.

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

[25] Junjun Zhang, Rosita Bajari, Dusan Andric, Francois Gerthoffert, Alexandru Lepsa, Hardeep Nahal-Bose, Lincoln D Stein, and Vincent Ferretti. The international cancer genome consortium data portal. *Nature biotechnology*, 37(4):367–369, 2019.

[26] Yara Skaf and Reinhard Laubenbacher. Topological data analysis in biomedicine: A review. *Journal of Biomedical Informatics*, 130:104082, 2022.

[27] Erik J Amézquita, Michelle Y Quigley, Tim Ophelders, Elizabeth Munch, and Daniel H Chitwood. The shape of things to come: Topological data analysis and biology, from molecules to organisms. *Developmental Dynamics*, 249(7):816–833, 2020.

[28] Allen Hatcher. *Algebraic topology*. 2005.

[29] James R Munkres. *Elements of algebraic topology*. CRC press, 2018.

[30] Robert Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[31] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4:667963, 2021.

[32] Gunnar Carlsson, Afra Zomorodian, Anne Collins, and Leonidas Guibas. Persistence barcodes for shapes. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 124–135, 2004.

[33] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[34] Herbert Edelsbrunner, John Harer, et al. Persistent homology-a survey. *Contemporary mathematics*, 453(26):257–282, 2008.

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

[36] Colleen Farrelly and Yae Ulrich Gaba. *The shape of data: Network Science, geometry-based machine learning, and topological data analysis in R*. No Starch Press, 2023.

[37] Paul Alexandroff. *Elementary concepts of topology*. Courier Corporation, 2012.

[38] Erin W Chambers, Vin De Silva, Jeff Erickson, and Robert Ghrist. Vietoris–rips complexes of planar point sets. *Discrete & Computational Geometry*, 44(1):75–90, 2010.

[39] Stefan Dantchev and Ioannis Ivrissimtzis. Efficient construction of the čech complex. *Computers & Graphics*, 36(6):708–713, 2012.

[40] Nataraj Akkiraju, Herbert Edelsbrunner, Michael Facello, Ping Fu, EP Mucke, and Carlos Varela. Alpha shapes: definition and software. In *Proceedings of the 1st international computational geometry software workshop*, volume 63, 1995.

[41] Vin De Silva and Gunnar E Carlsson. Topological estimation using witness complexes. In *PBG*, pages 157–166, 2004.

[42] Afra Zomorodian. Fast construction of the vietoris-rips complex. *Computers & Graphics*, 34(3):263–271, 2010.

[43] Ulrich Bauer. Ripser: efficient computation of vietoris–rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, 2021.

[44] Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & computational geometry*, 28:511–533, 2002.

[45] Herbert Edelsbrunner and John L Harer. *Computational topology: an introduction*. American Mathematical Society, 2022.

[46] Andrea Cerri, Barbara Di Fabio, Massimo Ferri, Patrizio Frosini, and Claudia Landi. Betti numbers in multidimensional persistent homology are stable functions. *Mathematical Methods in the Applied Sciences*, 36(12):1543–1557, 2013.

[47] Joel Friedman. Computing betti numbers via combinatorial laplacians. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of Computing*, pages 386–391, 1996.

[48] Seth Lloyd, Silvano Garnerone, and Paolo Zanardi. Quantum algorithms for topological and geometric analysis of data. *Nature communications*, 7(1):10138, 2016.

[49] Shashanka Ubaru, Ismail Yunus Akhalwaya, Mark S Squillante, Kenneth L Clarkson, and Lior Horesh. Quantum topological data analysis with linear depth and exponential speedup. *arXiv preprint arXiv:2108.02811*, 2021.

[50] Setup — TaDAsets 0.2.1 documentation.

[51] Peter Bubenik et al. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16(1):77–102, 2015.

[52] Peter Bubenik. The persistence landscape and some of its properties. In *Topological Data Analysis: The Abel Symposium 2018*, pages 97–117. Springer, 2020.

[53] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[54] Gerald N Wogan, Stephen S Hecht, James S Felton, Allan H Conney, and Lawrence A Loeb. Environmental and chemical carcinogenesis. In *Seminars in cancer biology*, volume 14, pages 473–486. Elsevier, 2004.

[55] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[56] John S Bertram. The molecular biology of cancer. *Molecular aspects of medicine*, 21(6):167–223, 2000.

[57] Saurav Kiri and Tyrone Ryba. Cancer, metastasis, and the epigenome. *Molecular Cancer*, 23(1):154, 2024.

[58] Maja Šutić, Ana Vukić, Jurica Baranašić, Asta Försti, Feđa Džubur, Miroslav Samaržija, Marko Jakopović, Luka Brčić, and Jelena Knežević. Diagnostic, predictive, and prognostic biomarkers in non-small cell lung cancer (nsclc) management. *Journal of personalized medicine*, 11(11):1102, 2021.

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

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

[61] Yousef Ahmed Fouad and Carmen Aanei. Revisiting the hallmarks of cancer. *American journal of cancer research*, 7(5):1016, 2017.

[62] Xiangming Guan. Cancer metastases: challenges and opportunities. *Acta pharmaceutica sinica B*, 5(5):402–418, 2015.

[63] Hanna Dillekås, Michael S Rogers, and Oddbjørn Straume. Are 90% of deaths from cancer caused by metastases? *Cancer medicine*, 8(12):5574–5576, 2019.

[64] Geoffrey M Cooper and Kenneth Adams. *The cell: a molecular approach*. Oxford University Press, 2022.

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

[66] 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.

[67] 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.

[68] 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.

[69] Manik Garg. *Prognostic biomarker discovery from omics data using machine learning approaches*. PhD thesis, University of Cambridge, 2022.

[70] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902, 2014.

[71] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17:1–19, 2016.

[72] Justin Matejka and George Fitzmaurice. Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 1290–1294, 2017.

[73] Comprehensive R Archive Network (CRAN). Datasets from the datasaurus dozen [r package datasaurus version 0.1.8], February 2024.

[74] Stephen M Stigler. Correlation and causation: a comment. *Perspectives in Biology and Medicine*, 48(1):88–S94, 2005.

[75] Ting Wang and Shiqiang Zhang. Study on linear correlation coefficient and nonlinear correlation coefficient in mathematical statistics. *Studies in Mathematical Sciences*, 3(1):58–63, 2011.

[76] Dhiren Ghosh and Andrew Vogt. Outliers: An evaluation of methodologies. In *Joint statistical meetings*, volume 12, pages 3455–3460, 2012.

[77] Guoqiang Zhong, Li-Na Wang, Xiao Ling, and Junyu Dong. An overview on data representation learning: From traditional feature learning to recent deep learning. *The Journal of Finance and Data Science*, 2(4):265–278, 2016.

[78] Rung-Ching Chen, Christine Dewi, Su-Wen Huang, and Rezzy Eko Caraka. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7(1):52, 2020.

[79] Monique F Kilkenny and Kerin M Robinson. Data quality:"garbage in–garbage out", 2018.

[80] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94–98, 2019.

[81] Rahul Shahane, Md Ismail, and CSR Prabhu. A survey on deep learning techniques for prognosis and diagnosis of cancer from microarray gene expression data. *Journal of computational and theoretical Nanoscience*, 16(12):5078–5088, 2019.

[82] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[83] Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

[84] Xiaojin Zhu. Persistent homology: An introduction and a new text representation for natural language processing. In *IJCAI*, number 2013, pages 1953–1959, 2013.

[85] Paul Michel, Abhilasha Ravichander, and Shruti Rijhwani. Does the geometry of word embeddings help document classification? a case study on persistent homology based representations. *arXiv preprint arXiv:1705.10900*, 2017.

[86] Hyekyoung Lee, Hyejin Kang, Moo K Chung, Bung-Nyun Kim, and Dong Soo Lee. Weighted functional brain network modeling via network filtration. In *NIPS workshop on algebraic topology and machine learning*, volume 3. Citeseer, 2012.

[87] Manish Saggar, Olaf Sporns, Javier Gonzalez-Castillo, Peter A Bandettini, Gunnar Carlsson, Gary Glover, and Allan L Reiss. Towards a new approach to reveal dynamical organization of the brain using topological data analysis. *Nature communications*, 9(1):1399, 2018.

[88] Lee M Seversky, Shelby Davis, and Matthew Berger. On time-series topological data analysis: New data and opportunities. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 59–67, 2016.

[89] Yuhei Umeda. Time series classification via topological data analysis. *Information and Media Technologies*, 12:228–239, 2017.

[90] Marian Gidea and Yuri Katz. Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical mechanics and its applications*, 491:820–834, 2018.

[91] Anubha Goel, Puneet Pasricha, and Aparna Mehra. Topological data analysis in investment decisions. *Expert Systems with Applications*, 147:113222, 2020.

[92] Lebohang Mashatola. *A Phenotype Prediction Framework for Classifying Colorectal Cancer Patients' Response to FOLFOX Treatment: An Integrated Approach*. PhD thesis, University of the Witwatersrand, Johannesburg, 2024.

[93] Sayan Mandal, Aldo Guzmán-Sáenz, Niina Haiminen, Saugata Basu, and Laxmi Parida. A topological data analysis approach on predicting phenotypes from gene expression data. In *International Conference on Algorithms for Computational Biology*, pages 178–187. Springer, 2020.

[94] Hosein Masoomy, Behrouz Askari, Samin Tajik, Abbas K Rizi, and G Reza Jafari. Topological analysis of interaction patterns in cancer-specific gene regulatory network: Persistent homology approach. *Scientific Reports*, 11(1):16414, 2021.

[95] Muhammad Sirajo Abdullahi, Apichat Suratanee, Rosario Michael Piro, and Kitiporn Plaimas. Persistent homology identifies pathways associated with hepatocellular carcinoma from peripheral blood samples. *Mathematics*, 12(5):725, 2024.

[96] Tamal K Dey, Sayan Mandal, and Soham Mukherjee. Gene expression data classification using topology and machine learning models. *BMC bioinformatics*, 22(Suppl 10):627, 2021.

[97] Lebohang Mashatola, Ismail Yunus Akhalwaya, and Stephanie Muller. Topological data analysis-deep learning framework for predicting cancer phenotypes. 2022.

[98] Renata Turkes, Guido F Montufar, and Nina Otter. On the effectiveness of persistent homology. *Advances in Neural Information Processing Systems*, 35:35432–35448, 2022.

[99] CNAM Oldenhuis, SF Oosting, JA Gietema, and EGE De Vries. Prognostic versus predictive value of biomarkers in oncology. *European journal of cancer*, 44(7):946–953, 2008.

[100] Virinder Kaur Sarhadi and Gemma Armengol. Molecular biomarkers in cancer. *Biomolecules*, 12(8):1021, 2022.

[101] Karla V Ballman. Biomarker: predictive or prognostic? *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 33(33):3968–3971, 2015.

[102] L Peter Fielding, Cecilia M Fenoglio-Preiser, and Laurence S Freedman. The future of prognostic factors in outcome prediction for patients with cancer. *Cancer*, 70(9):2367–2377, 1992.

[103] Victor M Corman, Olfert Landt, Marco Kaiser, Richard Molenkamp, Adam Meijer, Daniel KW Chu, Tobias Bleicker, Sebastian Brünink, Julia Schneider, Marie Luisa Schmidt, et al. Detection of 2019 novel coronavirus (2019-ncov) by real-time rt-pcr. *Eurosurveillance*, 25(3):2000045, 2020.

[104] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211, 2007.

[105] Nir Menachemi and Taleah H Collum. Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*, pages 47–55, 2011.

[106] Yahui Jiang, Meng Yang, Shuhao Wang, Xiangchun Li, and Yan Sun. Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer communications*, 40(4):154–166, 2020.

[107] Lili Qi, Kuiying Jiang, Fei-fei Zhao, Ping Ren, and Ling Wang. Identification of therapeutic targets and prognostic biomarkers in the siglec family of genes in tumor immune microenvironment of sarcoma. *Scientific Reports*, 14(1):577, 2024.

[108] Keith M Skubitz and David R D'Adamo. Sarcoma. In *Mayo Clinic Proceedings*, volume 82, pages 1409–1432. Elsevier, 2007.

[109] Anil K Rustgi and Hashem B El-Serag. Esophageal carcinoma. *New England Journal of Medicine*, 371(26):2499–2509, 2014.

[110] Ismail Essadi, Issam Lalya, and Hamid Mansouri. Esophageal carcinoma. *N Engl J Med*, 372(15):1470–1471, 2015.

[111] Rute Martins and Maria João Bugalho. Paragangliomas/pheochromocytomas: clinically oriented genetic testing. *International journal of endocrinology*, 2014(1):794187, 2014.

[112] Steven G Waguespack, T Rich, E Grubbs, AK Ying, ND Perrier, M Ayala-Ramirez, and C Jimenez. A current review of the etiology, diagnosis, and treatment of pediatric pheochromocytoma and paraganglioma. *The Journal of Clinical Endocrinology & Metabolism*, 95(5):2023–2037, 2010.

[113] Ryan D Rosen and Amit Sapra. Tnm classification. 2020.

[114] Ryan D Rosen and Amit Sapra. Tnm classification. In *StatPearls [Internet]*. StatPearls Publishing, 2023.

[115] James D Brierley, Elizabeth Van Eycken, Brian Rous, and Meredith Giuliani. *TNM classification of malignant tumours*. John wiley & sons, 2025.

[116] James D Brierley, Mary K Gospodarowicz, and Christian Wittekind. *TNM classification of malignant tumours*. John Wiley & Sons, 2017.

[117] S Edition, S Edge, D Byrd, et al. Ajcc cancer staging manual. *AJCC cancer staging manual*, 2017.

[118] Neha Tiwari, Alexander Gheldof, Marianthi Tatari, and Gerhard Christofori. Emt as the ultimate survival mechanism of cancer cells. In *Seminars in cancer biology*, volume 22, pages 194–207. Elsevier, 2012.

[119] Thomas Brabletz, Raghu Kalluri, M Angela Nieto, and Robert A Weinberg. Emt in cancer. *Nature Reviews Cancer*, 18(2):128–134, 2018.

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

[121] Gerald McMahon. Vegf receptor signaling in tumor angiogenesis. *The oncologist*, 5(S1):3–10, 2000.

[122] Peter Vaupel and Arnulf Mayer. Hypoxia in cancer: significance and impact on clinical outcome. *Cancer and Metastasis Reviews*, 26:225–239, 2007.

[123] M Christiane Brahimi-Horn, Johanna Chiche, and Jacques Pouysségur. Hypoxia and cancer. *Journal of molecular medicine*, 85:1301–1307, 2007.

[124] William R Wilson and Michael P Hay. Targeting hypoxia in cancer therapy. *Nature Reviews Cancer*, 11(6):393–410, 2011.

[125] Claudio R Santos and Almut Schulze. Lipid metabolism in cancer. *The FEBS journal*, 279(15):2610–2623, 2012.

[126] Charlene Brault and Almut Schulze. The role of glucose and lipid metabolism in growth and survival of cancer cells. *Metabolism in Cancer*, pages 1–22, 2016.

[127] Gurjeet Singh, Facundo Mémoli, Gunnar E Carlsson, et al. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG@ Eurographics*, 2:091–100, 2007.

[128] 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*, 108(17):7265–7270, 2011.

[129] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9:1–13, 2008.

[130] R Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

[131] Michael I Love, John B Hogenesch, and Rafael A Irizarry. Modeling of rna-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature biotechnology*, 34(12):1287–1291, 2016.

[132] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

[133] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15:1–21, 2014.

[134] Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics*, 13:1–21, 2012.

[135] Gwenaëlle G Lemoine, Marie-Pier Scott-Boyer, Bathilde Ambroise, Olivier Périn, and Arnaud Droit. Gwena: gene co-expression networks analysis and extended modules characterization in a single bioconductor package. *BMC bioinformatics*, 22(1):267, 2021.

[136] Philip Sedgwick. Pearson's correlation coefficient. *Bmj*, 345, 2012.

[137] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[138] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. 2007.

[139] Carlos Ramos-Carreño and José L. Torrecilla. dcor: Distance correlation and energy statistics in Python. *SoftwareX*, 22, 2 2023.

[140] Andy M Yip and Steve Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics*, 8:1–14, 2007.

[141] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[142] André Voigt and Eivind Almaas. Assessment of weighted topological overlap (wto) to improve fidelity of gene co-expression networks. *BMC bioinformatics*, 20:1–11, 2019.

[143] Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. The gudhi library: Simplicial complexes and persistent homology. In *Mathematical Software–ICMS 2014: 4th International Congress, Seoul, South Korea, August 5-9, 2014. Proceedings 4*, pages 167–174. Springer, 2014.

[144] Marc Glisse and Siddharth Pritam. Swap, shift and trim to edge collapse a filtration. *arXiv preprint arXiv:2203.07022*, 2022.

[145] Davide Gurnari, Aldo Guzmán-Sáenz, Filippo Utro, Aritra Bose, Saugata Basu, and Laxmi Parida. Probing omics data via harmonic persistent homology. *arXiv preprint arXiv:2311.06357*, 2023.

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

[147] Corinna Cortes. Support-vector networks. *Machine Learning*, 1995.

[148] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[149] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[150] Aric Hagberg, Pieter J Swart, and Daniel A Schult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), 2008.

[151] Paul Barrett, John Hunter, J Todd Miller, J-C Hsu, and Perry Greenfield. matplotlib–a portable python plotting package. In *Astronomical data analysis software and systems XIV*, volume 347, page 91, 2005.

[152] Giancarlo Perrone, Jose Unpingco, and Haw-minn Lu. Network visualizations with pyvis and visjs. *arXiv preprint arXiv:2006.04951*, 2020.

[153] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[154] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[155] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[156] Panneerselvam Jayabal, Fuchun Zhou, Xiufen Lei, Xiuye Ma, Barron Blackman, Susan T Weintraub, Peter J Houghton, and Yuzuru Shiio. Nell2-cdc42 signaling regulates baf complexes and ewing sarcoma cell growth. *Cell reports*, 36(1), 2021.

[157] Awf A Al-Khan, Noora R Al Balushi, Samantha J Richardson, and Janine A Danks. Roles of parathyroid hormone-related protein (pthrp) and its receptor (pthr1) in normal and tumor tissues: Focus on their roles in osteosarcoma. *Frontiers in veterinary science*, 8:637614, 2021.

[158] Awf A Al-Khan, Judith S Nimmo, Mourad Tayebi, Stewart D Ryan, James O Simcock, Raboola Tarzi, Charles A Kuntz, Eman S Saad, Michael J Day, Samantha J Richardson, et al. Parathyroid hormone receptor 1 (pthr1) is a prognostic indicator in canine osteosarcoma. *Scientific reports*, 10(1):1564, 2020.

[159] Jifeng Wen, Yong Qin, Chao Li, Xiankui Dai, Tong Wu, and Wenzhe Yin. Mangiferin suppresses human metastatic osteosarcoma cell growth by down-regulating the expression of metalloproteinases-1/2 and parathyroid hormone receptor 1. *AMB Express*, 10:1–10, 2020.

[160] Shenglong Li, Yi Pei, Wei Wang, Fei Liu, Ke Zheng, and Xiaojing Zhang. Quercetin suppresses the proliferation and metastasis of metastatic osteosarcoma cells by inhibiting parathyroid hormone receptor 1. *Biomedicine & Pharmacotherapy*, 114:108839, 2019.

[161] Hitoshi Kimura, Hiroyuki Kato, Ahmad Faried, Makoto Sohda, Masanobu Nakajima, Yasuyuki Fukai, Tatsuya Miyazaki, Norihiro Masuda, Minoru Fukuchi, and Hiroyuki Kuwano. Prognostic significance of epcam expression in human esophageal cancer. *International journal of oncology*, 30(1):171–179, 2007.

[162] Tatsuo Matsuda, Hiroya Takeuchi, Sachiko Matsuda, Kunihiko Hiraiwa, Taku Miyasho, Minoru Okamoto, Kazufumi Kawasako, Rieko Nakamura, Tsunehiro Takahashi, Norihito Wada, et al. Epcam, a potential therapeutic target for esophageal squamous cell carcinoma. *Annals of Surgical Oncology*, 21:356–364, 2014.

[163] Nikolas H Stoecklein, Annika Siegmund, Peter Scheunemann, Andreas M Luebke, Andreas Erbersdobler, Pablo E Verde, Claus F Eisenberger, Matthias Peiper, Alexander Rehders, Jan Schulte am Esch, et al. Ep-cam expression in squamous cell carcinoma of the esophagus: a potential therapeutic target and prognostic marker. *BMC cancer*, 6:1–8, 2006.

[164] Raghibul Hasan, Rinu Sharma, Anoop Saraya, Tushar K Chattopadhyay, Siddartha DattaGupta, Paul G Walfish, Shyam S Chauhan, and Ranju Ralhan. Mitogen activated protein kinase kinase kinase 3 (map3k3/mekk3) overexpression is an early event in esophageal tumorigenesis and is a predictor of poor disease prognosis. *BMC cancer*, 14:1–7, 2014.

[165] Annica Wilzén, Anna Rehammar, Andreas Muth, Ola Nilsson, Tajana Tešan Tomić, Bo Wängberg, Erik Kristiansson, and Frida Abel. Malignant pheochromocytomas/paragangliomas harbor mutations in transport and cell adhesion genes. *International journal of cancer*, 138(9):2201–2211, 2016.

[166] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[167] Yuhan Sun, Md Abdullah Al Kamran Khan, Stefano Mangiola, and Alexander David Barrow. Il17rb and il17rel expression are associated with improved prognosis in hpv-infected head and neck squamous cell carcinomas. *Pathogens*, 12(4):572, 2023.

[168] Bray Denard, Sharon Jiang, Yan Peng, and Jin Ye. Creb3l1 as a potential biomarker predicting response of triple negative breast cancer to doxorubicin-based chemotherapy. *BMC cancer*, 18:1–7, 2018.

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

[170] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[171] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[172] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655, 2018.

[173] Jay Gambetta. Quantum-centric supercomputing: The next wave of computing. *IBM Research Blog*, 2022.

[174] Charles Q Choi. Ibm's quantum leap: The company will take quantum tech past the 1,000-qubit mark in 2023. *IEEE Spectrum*, 60(1):46–47, 2023.

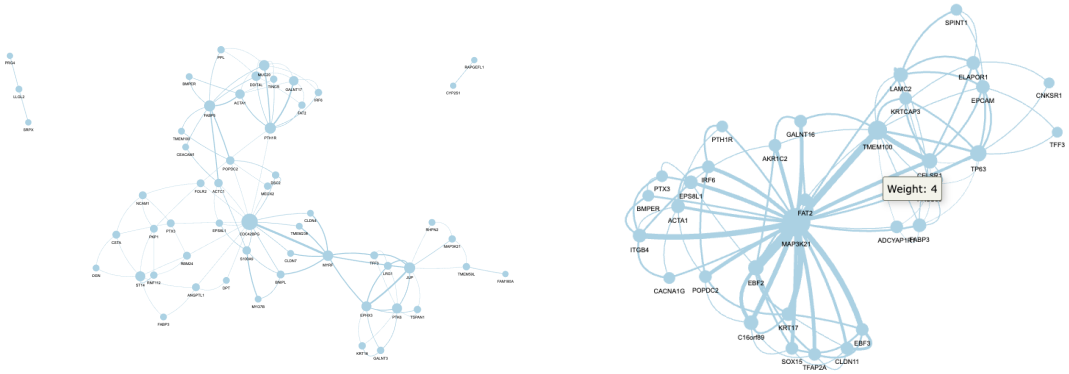
[175] Muhammad AbuGhanem. Ibm quantum computers: Evolution, performance, and future directions. *arXiv preprint arXiv:2410.00916*, 2024.

[176] Saugata Basu and Nathanael Cox. Harmonic persistent homology. *SIAM Journal on Applied Algebra and Geometry*, 8(1):189–224, 2024.

Appendix A

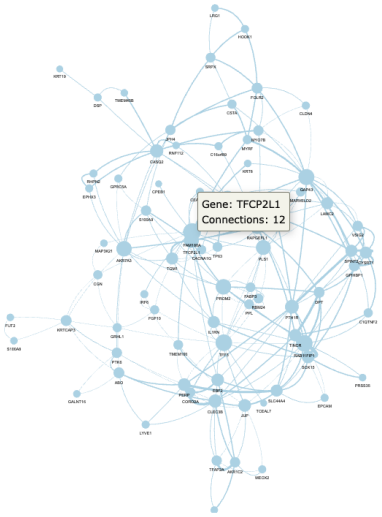
Supplementary Figures

A.1 Web-based Interactive WGTDA Networks



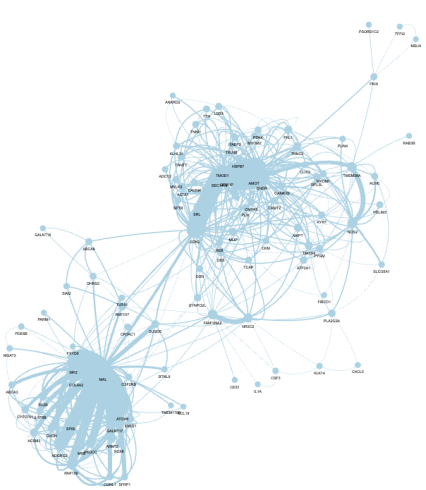
(a) SARC

(b) ESCA

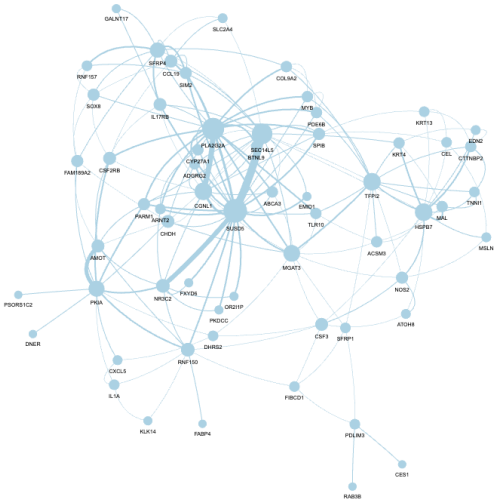


(c) PCPG

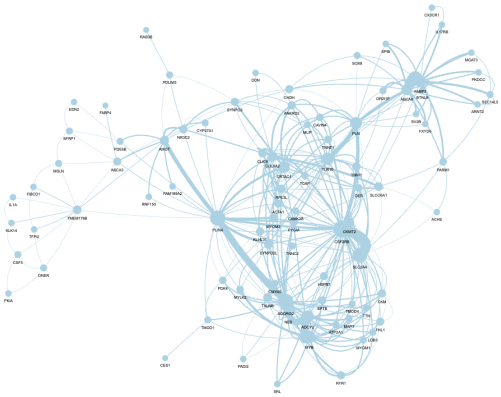
Fig. A.1 WGTDA Web-based Network illustrating topological features visualized as a network for each cancer type: (a) SARC, (b) ESCA, (c) PCPG.



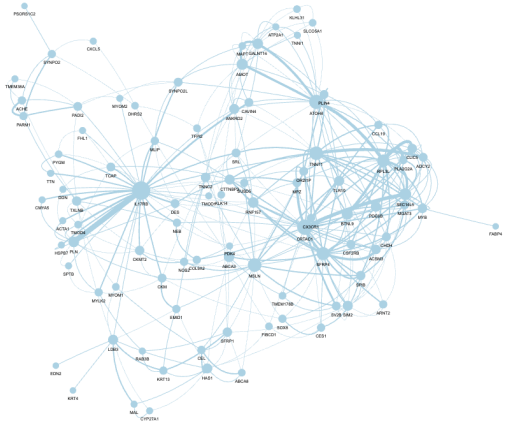
(a) Stage 1



(b) Stage 2

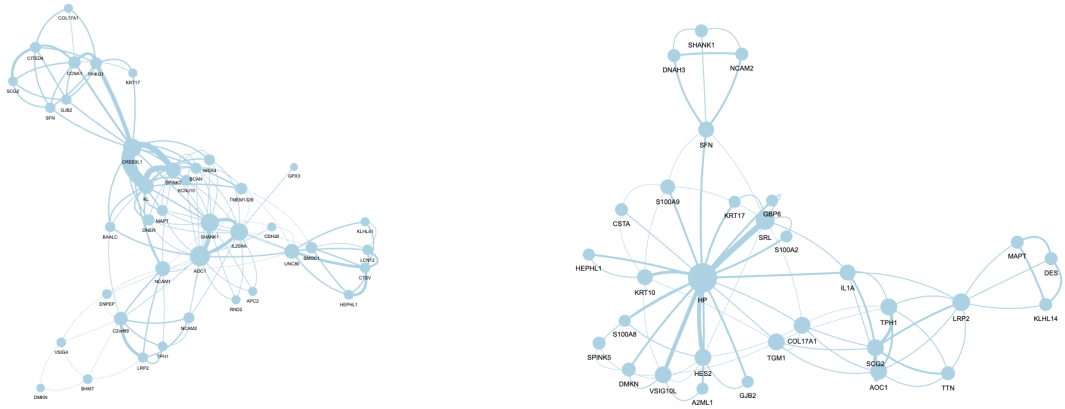


(c) Stage 3



(d) Stage 4

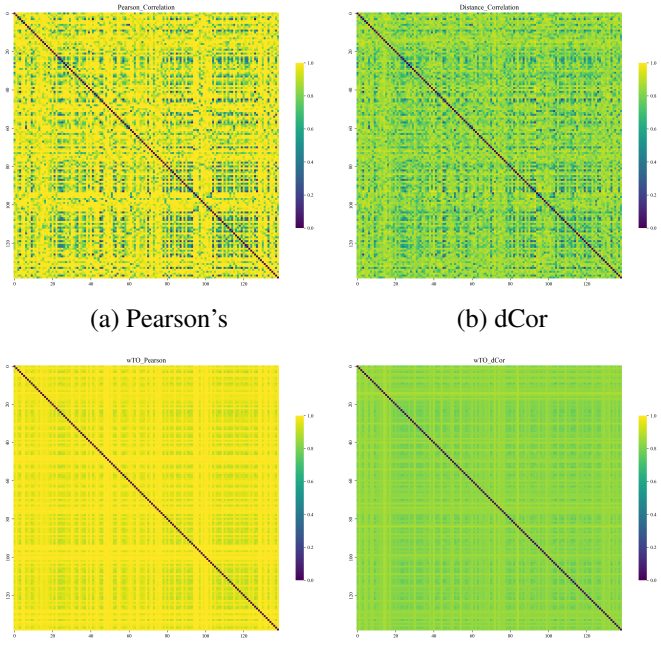
Fig. A.2 WGTDA Web-based Network illustrating topological features visualized as a network for each cancer stage (HNSC): (a) Stage I, (b) Stage II, (c) Stage III (d) Stage IV.



(a) Resistant (b) Sensitive

Fig. A.3 WGTDA Web-based Network illustrating topological features visualized as a network for resistant and sensitive patients: (a) Resistant, (b) Sensitive

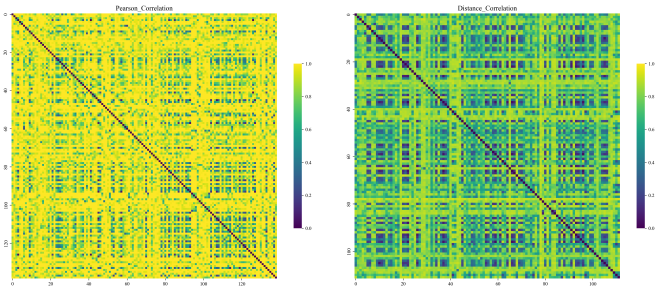
A.2 Co-expression Analysis



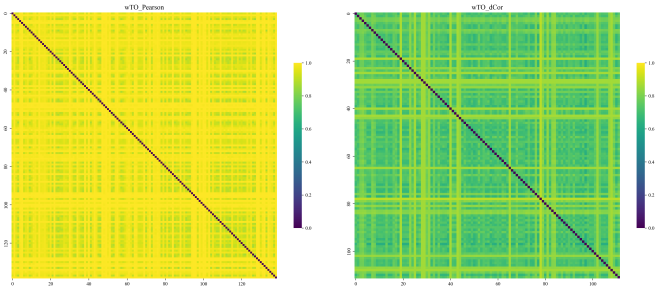
(a) Pearson's (b) dCor

(c) wTO-Pearson's (d) wTO-dCor

Fig. A.4 Heatmap Comparison of Co-expression Measures for SARC: (a) Pearson's, (b) Distance Correlation (dCor), (c) Weighted Topological Overlap (wTO) with Pearson's, and (d)wTO with dCor.

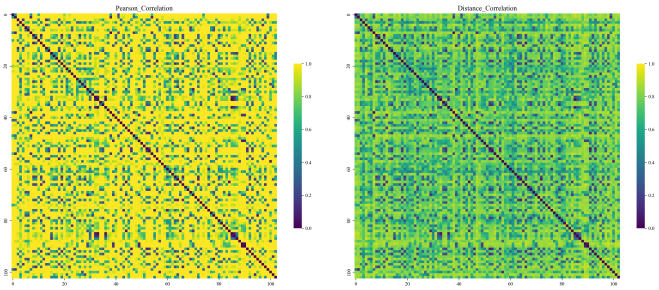


(a) Pearson's (b) dCor

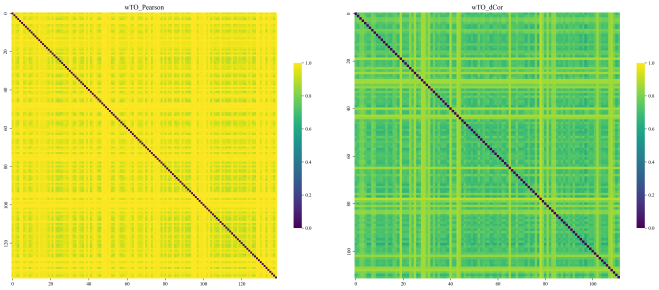


(c) wTO-Pearson's (d) wTO-dCor

Fig. A.5 Heatmap Comparison of Co-expression Measures for Stage IV (HNSC): (a) Pearson's, (b) Distance Correlation (dCor), (c) wTO with Pearson's, and (d) wTO with dCor.



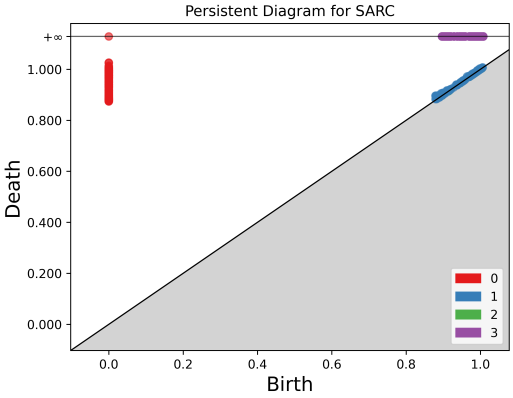
(a) Pearson's (b) dCor



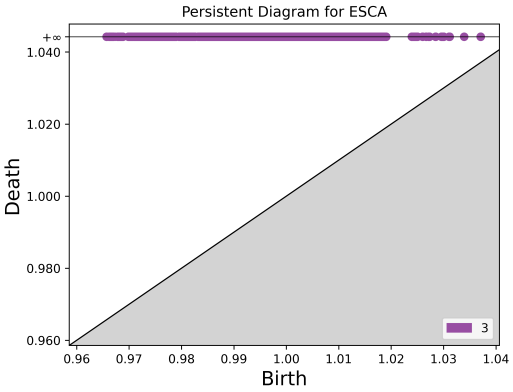
(c) wTO-Pearson's (d) wTO-dCor

Fig. A.6 Heatmap Comparison of Co-expression Measures for Resistance Treatment Response: (a) Pearson's, (b) Distance Correlation (dCor), (c) wTO with Pearson's, and (d) wTO with dCor.

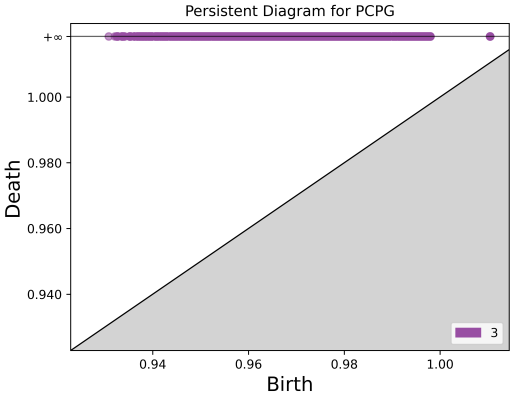
A.3 Persistent Diagrams



(a) SARC

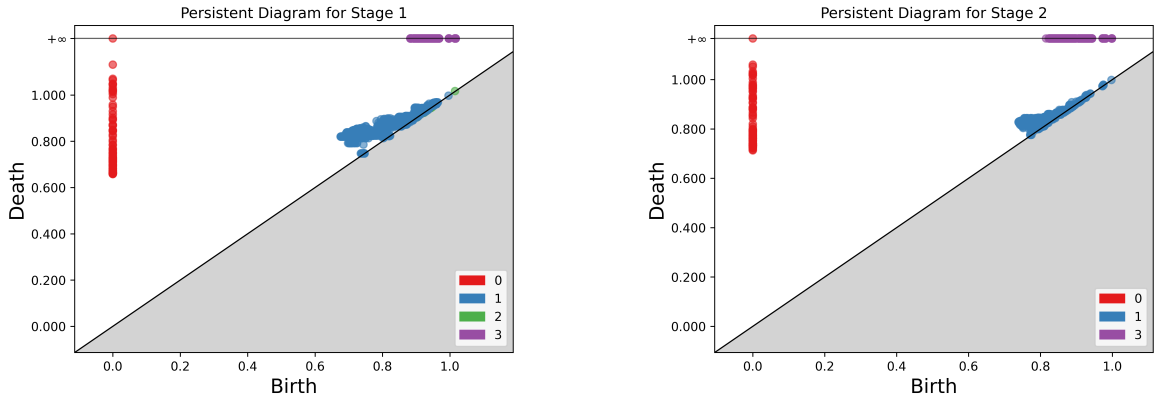


(b) ESCA



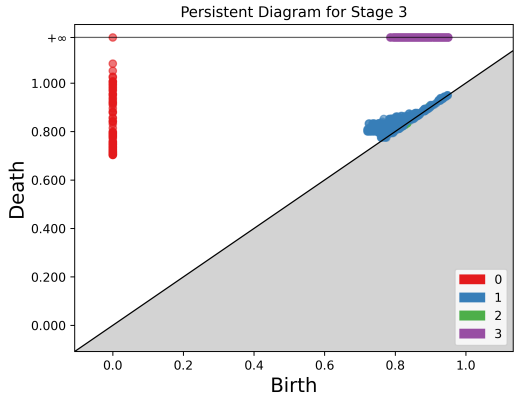
(c) PCPG

Fig. A.7 Persistent diagrams illustrating topological features for each of the three cancer types: (a) SARC, (b) ESCA, (c) PCPG. Points are color-coded by Betti number: red for β_0, blue for β_1, green for β_2, and purple for β_3.

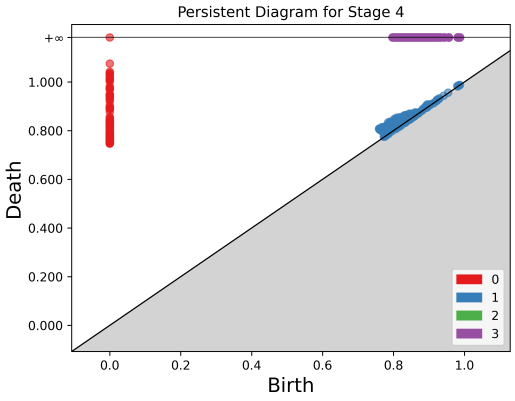


(a) Stage 1

(b) Stage 2

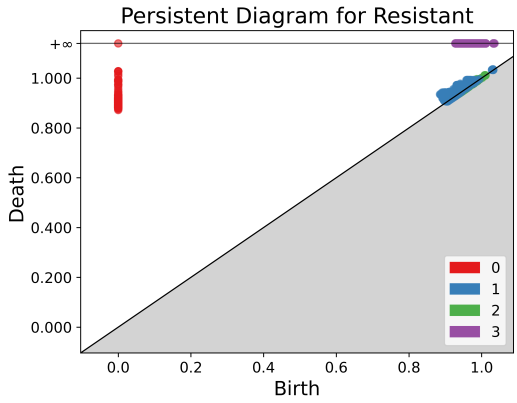


(c) Stage 3

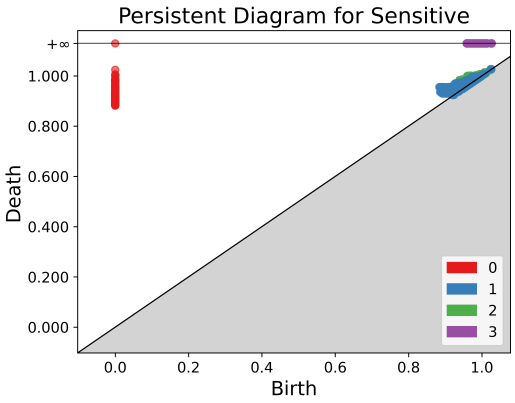


(d) Stage 4

Fig. A.8 Persistent diagrams illustrating topological features for each of the four HNSC cancer stages: (a) Stage I, (b) Stage II, (c) Stage III, and (d) Stage IV. Points are color-coded by Betti number: red for β_0, blue for β_1, green for β_2, and purple for β_3.



(a) Resistant



(b) Sensitive

Fig. A.9 Persistent diagrams illustrating topological features (a) Resistant (b) Sensitive. Points are color-coded by Betti number: red for β_0, blue for β_1, green for β_2, and purple for β_3.