Politecnico di Torino

Master's Degree in Data Science and Engineering

Master's Degree Thesis

Investigating LIONESS-Derived Gene Regulatory Network and Advanced Graph Neural Network for Leukemia Subtype Classification and Cross-Cancer Disease Association Analysis

Supervisors

Prof. Roberta Bardini
Prof. Stefano Di Carlo
Prof. Alessandro Savino
Dr. Lorenzo Martini
Dr. Riccardo Smeriglio

Candidate

Cesar Augusto Seminario Yrigoyen

March 2026

Acknowledgements

I would like to thank my wife and my two daughters for supporting and enduring me throughout these years. The time I devoted to this work, partially taken away from my family, represents a measure of the sacrifice I required to complete this journey. I hope this work will show my daughters that achieving meaningful goals requires dedication, perseverance, and a willingness to make sacrifices. I sincerely hope I have been able to set a positive example for them.

I would also like to express my sincere gratitude to my supervisors, Dr. Riccardo Smeriglio, Dr. Lorenzo Martini and Prof. Roberta Bardini for their constant support and close guidance throughout this work. They understood my moments of difficulty and provided invaluable mentorship and encouragement. Their suggestions helped overcome several obstacles and significantly improved the quality and impact of this research.

Finally, I would like to thank my mother, who has always been by my side, my brothers and sisters, Giuseppe and my close friends.

AI Assistance Statement

AI-based language tools were used to support grammar and stylistic refinement of the manuscript. All scientific content and analyses presented in this thesis are the original work of the author.

Abstract

Cancer classification has progressively shifted from histopathological assessment to molecular profiling, particularly RNA sequencing (RNAseq) gene expression and, more recently, multi-omics integration approaches. Most methods rely on static gene features or predefined interaction networks and rarely incorporate gene regulatory structure. In addition, cross-cancer disease association remain largely disconnected from individualized gene regulatory modeling. This thesis addresses this gap by integrating patient specific Gene Regulatory Network (GRN)s -inferred via LIONESS- with Graph Neural Network (GNN) models, exploring an under-investigated space at the intersection of multi-omics integration, network biology, and cross-disease cancer modeling.

The origin datasets are retrieved from the Genomic Data Commons (GDC) portal to construct a large, curated RNAseq cohort comprising leukemia, breast cancer and normal controls.

Main approach is built on top of the leukemia dataset, two modelling paradigms are explored. First, GEX-based models operate directly on the gene expression matrix using variance-driven feature selection and conventional Machine Learning (ML) or shallow neural networks. These models achieve high predictive performance and provide a baseline for comparison. However, their formulation remains fundamentally feature-centric and does not explicitly encode regulatory structure.

The core contribution of this work lies in the development of GRN-based models. GRNs are inferred using PANDA, which integrates transcription factor (TF)-gene priors with protein–protein interaction information to estimate regulatory edge strengths. To capture inter-patient heterogeneity, LIONESS is applied to derive single-sample regulatory networks, enabling the reconstruction of patient-specific regulatory landscapes. These networks encode TF-gene interactions and global regulatory topology, reflecting the systems-level organisation of transcriptional control.

Patient specific networks are subsequently transformed into graph representations suitable for graph-level classification. Advanced GNN architectures including GCN, GAT, GraphSAGE, GINE, and GRNFormer are evaluated to model high-dimensional regulatory graphs containing millions of edges. Although predictive performance is lower than purely expression-based models, the GRN-based framework provides a biologically grounded representation of disease mechanisms and enables regulatory level interpretability. SHAP and GNNExplainer are employed to identify key regulatory genes

and discriminative subnetworks, which are validated against literature reported differentially expressed genes and known oncogenic pathways.

To assess the generalizability of the proposed framework, RNAseq data from breast cancer (TCGA-BRCA), are integrated with the leukemia cohort to construct a joint multi-tumor dataset. The entire modelling pipeline is re-applied without structural modification to develop a unified classifier that distinguishes tumour versus normal samples across tissues, following a disease-association paradigm. This cross-cancer evaluation demonstrates the generalizability of the proposed framework beyond leukaemia subtype classification, supporting its applicability to heterogeneous multi tissue oncological settings.

Contents

List of Acronyms

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
GNN	Graph Neural Network
GRN	Gene Regulatory Network
SVM	Support Vector Machine
RF	Random Forest
RNAseq	RNA sequencing
AML	Acute myeloid leukemia
ALL	Acute Lymphoblastic Leukemia

Chapter 1

Introduction

1.1 Background and Motivation

High-throughput RNA sequencing (RNAseq) has become a foundational technology in modern genomics, enabling systematic and quantitative profiling of gene expression across diverse biological conditions and disease states. Bulk RNAseq captures averaged transcriptional activity across heterogeneous cell populations, while single-cell RNA sequencing (scRNAseq) resolves expression variability at cellular resolution, revealing transcriptional heterogeneity that remains obscured in bulk measurements [16, 46]. The widespread availability of harmonised transcriptomic repositories, such as ThCancer Genome Atlas (TCGA) and the Genomic Data Commons (GDC) [66], has substantially accelerated integrative cancer research by providing large-scale RNAseq datasets accompanied by detailed clinical annotations.

RNAseq data are routinely employed for supervised classification tasks, including disease diagnosis, cancer subtype stratification, prognosis prediction, and treatment-response assessment. Transcriptomic profiling has demonstrated that gene expression signatures can reliably distinguish between leukemia subtypes and other cancer phenotypes, supporting its use in molecular diagnostics and precision medicine [54, 86].

Traditional ML approaches treat each patient sample as a high-dimensional feature vector, where genes are assumed to be independent predictors. Algorithms such as Random Forest (RF), Support Vector Machine (SVM), and multilayer perceptrons are widely applied to gene expression datasets and frequently achieve high predictive performance in biomedical classification problems [31, 96].

However, the assumption of gene independence conflicts with established principles of systems biology and network medicine. Biological phenotypes

emerge from coordinated perturbations of interacting molecular components rather than isolated gene alterations. As emphasised in network medicine, disease mechanisms are better understood as dysregulations of molecular interaction networks [11]. In this view, cancer represents a rewiring of regulatory and signalling networks rather than merely a change in individual gene expression levels.

To incorporate such relational structure, graph-based representations of transcriptomic data have gained increasing attention. In this framework, genes are represented as nodes and edges encode functional, regulatory, or statistical dependencies. GRNs provide a mechanistic representation of transcriptional control by modelling how transcription factors influence downstream gene expression under specific biological contexts. Among GRN inference approaches, PANDA (Passing Attributes between Networks for Data Assimilation) integrates gene expression with transcription factor binding priors and protein-protein interaction networks to infer condition specific regulatory networks [33].

While PANDA produces an aggregate network representative of a cohort, it does not explicitly capture inter-patient heterogeneity-one of the defining characteristics of cancer, particularly in haematological malignancies. The LIONESS framework (Linear Interpolation to Obtain Network Estimates for Single Samples) addresses this limitation by enabling the extraction of patient-specific regulatory networks from an aggregate reconstruction [48]. Through a linear interpolation strategy, LIONESS estimates the contribution of each individual sample to the global network, thereby generating a distinct regulatory graph for every patient.

Once transcriptomic data are structured as graphs, GNNs provide a principled learning framework capable of jointly leveraging node features and graph topology. By propagating information across neighbouring nodes, GNNs model higher order dependencies and relational patterns that are inaccessible to vector based approaches. Recent work demonstrates that incorporating biological interaction networks into GNN-based phenotype prediction can improve modelling of complex interactions and reveal epistatic or regulatory patterns not captured by linear models [81].

Despite these advances, two major challenges remain. First, predictive performance alone is insufficient in biomedical contexts, where model outputs must be biologically interpretable and clinically trustworthy. Second, increasing structural complexity such as modelling millions of regulatory edges does not guarantee improved predictive accuracy and introduces significant computational challenges. These considerations motivate a systematic investigation of graph-based transcriptomic modelling and explainability in the context of inter-patient heterogeneity in leukaemia and beyond.

1.2 Problem Statement and Research Questions

Most RNAseq classification pipelines rely on feature-based representations that ignore regulatory dependencies among genes. Although such models can achieve near-perfect predictive performance, they offer limited mechanistic insight into the regulatory processes underlying disease phenotypes.

Conversely, regulatory-network-based approaches introduce biologically meaningful structure but substantially increase modelling and computational complexity. In particular:

- PANDA derived regulatory networks contain millions of transcription factor gene interactions.
- LIONESS generates a distinct regulatory graph for each patient, resulting in extremely large graph datasets.
- GNNs applied to such networks must address scalability, memory constraints, and optimisation stability.
- Explainability methods for GNNs remain underexplored in transcriptomic applications and are rarely benchmarked systematically.

Furthermore, the literature often fails to clearly distinguish between biologically grounded regulatory networks and statistically constructed association graphs (e.g., patientâpatient or geneâgene Pearson correlation networks), despite their fundamentally different interpretative implications. Correlation-based approaches capture statistical dependencies between gene expression profiles but do not explicitly encode mechanistic regulatory interactions or causal transcriptional control [48, 77].

Therefore, the **central problem addressed in this thesis** is:

> How can LIONESS-derived patient-specific GNNs be effectively integrated within explainable GNN architectures to model inter-patient heterogeneity in leukaemia, and how does this structured approach compare to classical gene expression-based models in terms of predictive performance, interpretability, scalability, and biological coherence?

Additionally, an open methodological question concerns generalisability:

> Can a regulatory-network-based pipeline developed for leukaemia subtype classification generalise across distinct tumour types under a disease-association paradigm?

The **primary objective** of this thesis is to design, implement, and systematically evaluate a structured modelling pipeline that integrates:

- RNAseq expression matrices,
- PANDA-based GRN inference,
- LIONESS single sample network reconstruction,
- GNN classification,
- and explainability techniques for biological validation.

This objective is operationalised through the following **research questions**:

1. **RQ1:** How does the predictive performance of classical GEX-based models compare with GRN-based graph models in leukaemia subtype classification (AML vs Acute Lymphoblastic Leukemia (ALL) vs normal)?
2. **RQ2:** Does incorporating regulatory structure via PANDA and LIONESS enhance biological interpretability, even when predictive accuracy is lower?
3. **RQ3:** Which GNN architectures are most suitable for large-scale regulatory graphs in terms of scalability and optimisation stability?
4. **RQ4:** How stable and biologically coherent are explanations produced by SHAP and GNNExplainer in both feature-based and graph-based settings?
5. **RQ5:** Can the entire modelling pipeline generalise across tumour types, enabling a unified disease-association classifier distinguishing normal versus tumour samples in a multi-tissue setting (leukaemia + breast cancer)?

This thesis provides methodological, computational, and empirical contributions. First, a curated large-scale leukaemia cohort was constructed from the GDC portal, comprising over 2,500 RNAseq samples, including Acute myeloid leukemia (AML), ALL, and normal tissues. Only primary-diagnosis tumours were retained to avoid recurrence-related confounding effects. Expression values were quantified as FPKM and harmonised with clinical annotations.

Second, two complementary modelling branches were developed:

1. **GEX based models**, operating directly on gene expression matrices using dimensionality reduction and classical ML algorithms (RF, SVM, multilayer perceptron, feedforward neural networks), as well as patient similarity graphs analysed via GNN architectures (GCN, GAT, GraphSAGE).

2. **GRN based models**, incorporating regulatory information through PANDA and LIONESS. Patient specific regulatory networks containing approximately nine million transcription factor gene interactions were transformed into graph structures suitable for GNN classification.

Third, scalable data engineering solutions were implemented to handle large single-sample regulatory networks. Efficient storage and lazy-loading strategies reduced memory footprint by an order of magnitude and significantly shortened training time.

Fourth, a systematic explainability analysis was conducted using SHAP, saliency methods, integrated gradients, and GNNExplainer to identify discriminative genes and subnetworks. These were validated against literature-reported differentially expressed genes and pathway signatures.

Fifth, to assess robustness and generalisability, the entire pipeline was extended to an independent tumour cohort. RNAseq data from the TCGA-BRCA project were integrated with the leukaemia dataset to construct a joint multi-tumour cohort. A unified binary classifier (normal vs tumour) was trained without structural modification of the pipeline, following a disease-association paradigm. This transversal evaluation demonstrates that the proposed framework is not restricted to leukaemia subtype discrimination but can capture shared oncogenic regulatory patterns across distinct tissues.

1.3 Structure of the Thesis

The remainder of this thesis is organised as follows:

- **Chapter 2 - Biological and Computational Foundations**

  Reviews RNAseq technologies, GNN inference methods (including PANDA and LIONESS), network medicine, and GNNs in biomedical applications.

- **Chapter 3 - Dataset and Pre-processing**

  Describes dataset curation, preprocessing for RNAseq and GRN reconstruction along with graph construction.

- **Chapter 4 - Methods and Modeling Framework**

  Describes classical ML and GNN architectures, hyperparameter optimisation, and explainability methodologies.

- **Chapter 5 - Results and Discussion**

  Presents comparative evaluation of GEX-based and GRN-based models, scalability analysis, and cross-cancer disease-association experiments. Interprets explainability results in the context of known AML and ALL molecular mechanisms and discusses strengths and limitations of regulatory graph modelling

- **Chapter 6 - Conclusions and Future Direcions**

  Summarises findings and outlines future directions for scalable and interpretable graph-based modelling of multi-omic data.

Chapter 2

Biological and Computational Foundations

2.1 From Cellular Information Flow to Transcriptomic Measurement

2.1.1 The Cell as an Information Processing System

The eukaryotic cell can be conceptualized as a dynamic and hierarchical information processing system in which molecular interactions govern functional behavior. Cellular identity and phenotype are determined by the controlled flow of biological information from the genome to functional molecular effectors. This flow is not static; rather, it is continuously modulated by regulatory mechanisms that enable adaptation, differentiation, and response to environmental stimuli.

From DNA to Functional Output Deoxyribonucleic acid (DNA) constitutes the stable repository of genetic information. Organized within chromatin in the cell nucleus, DNA encodes the complete set of genes required for cellular function. However, the genome itself does not directly determine phenotype. Instead, phenotype emerges from the selective activation and repression of genes through tightly regulated transcriptional processes.

Transcription is the biochemical mechanism by which specific DNA sequences are transcribed into ribonucleic acid (RNA) molecules. Among these, messenger RNA (mRNA) plays a central role by serving as an intermediary between genomic information and protein synthesis. The abundance of mRNA transcripts reflects the transcriptional state of the cell and provides a quantitative measure of gene activity.

Importantly, transcription is not constitutive. Only a subset of genes is expressed at any given time, and expression levels vary according to cell type, developmental stage, and environmental context. Thus, gene expression profiles represent the operational state of the cell rather than merely its genetic potential.

RNA molecules undergo post-transcriptional processing, including splicing, capping, and polyadenylation, before being translated into proteins. Proteins execute structural, enzymatic, and regulatory functions that collectively determine cellular behavior. Consequently, cellular phenotype arises as an emergent property of coordinated gene expression programs.

This systems-level perspective aligns with the framework of network medicine, in which diseases are interpreted as perturbations of interconnected molecular systems rather than isolated gene defects [11].

Gene Expression as a Proxy for Regulatory Activity Because transcription represents the primary interface between the genome and cellular function, transcriptomic measurements provide indirect access to underlying regulatory mechanisms. RNAseq technologies enable high-throughput quantification of mRNA abundance, thereby offering a snapshot of the transcriptional landscape of a cell or tissue.

Beyond measuring individual genes, transcriptomic datasets also capture relationships among genes. In particular, genes participating in the same biological pathways or controlled by common transcription factors often exhibit coordinated expression patterns across samples. These statistical dependencies provide valuable information about potential regulatory relationships and are widely exploited to reconstruct GRNs, which aim to model how transcription factors influence downstream gene expression [16].

Thus, transcriptomic profiling does not merely quantify gene abundance; it encodes relational information reflective of the regulatory architecture governing cellular function.

Transcription Factors as Master Regulators of Cellular Identity Central to transcriptional regulation are transcription factors (TFs), proteins that bind to specific DNA sequences and modulate the transcriptional activity of target genes [24, 74]. TFs recognize short regulatory motifs located in promoter or enhancer regions and can function either as activators or repressors depending on cellular context.

Transcription factors operate combinatorially, forming regulatory circuits that coordinate the expression of large gene sets [74]. Through these interactions, TFs define cell identity, lineage commitment, and response to signaling

cues. Dysregulation of TF activity can therefore result in widespread transcriptional reprogramming and contribute to disease pathogenesis.

In hematopoiesis, lineage specification is governed by tightly regulated TF networks. Disruption of these networks is a hallmark of leukemia. In acute myeloid leukemia (AML), transcription factors such as *RUNX1*, *CEBPA*, and *MYB* play critical roles in myeloid differentiation, and their mutation or aberrant regulation leads to impaired maturation and uncontrolled proliferation [67, 74, 83]. Similarly, in B-cell acute lymphoblastic leukemia (B-ALL), TFs such as *PAX5* and *TCF3* (E2A) regulate B-cell lineage identity and are frequently altered in leukemogenesis, [20, 88].

Beyond hematologic malignancies, oncogenic transcription factors such as *MYC* are broadly implicated across multiple cancer types. *MYC* functions as a global transcriptional amplifier, modulating genes involved in cell cycle progression, metabolism, and apoptosis. Its deregulation induces extensive transcriptional rewiring and supports tumor maintenance [24].

These examples illustrate that transcription factors act as master regulators of cellular phenotype. Their dysregulation does not merely affect isolated target genes but reshapes entire regulatory programs. Consequently, modelling TF–gene interactions at the network level provides a biologically grounded representation of disease mechanisms.

From Molecular Interactions to Phenotypic Heterogeneity Cellular phenotype ultimately emerges from the integrated output of regulatory networks composed of transcription factors, target genes, protein–protein interactions, and signaling pathways. In cancer, somatic mutations, chromosomal rearrangements, and epigenetic alterations disrupt these networks, leading to regulatory rewiring. Such rewiring generates phenotypic heterogeneity both within and across patients. [11]

In leukemia, this heterogeneity manifests as distinct molecular subtypes characterized by divergent transcriptional programs . While traditional transcriptomic classification approaches often treat genes as independent features, biological reality suggests that genes operate within coordinated regulatory systems. [54]

Therefore, to accurately model disease mechanisms and inter-patient variability, it is necessary to move beyond flat gene expression matrices and adopt network-based representations that explicitly encode regulatory interactions. [11, 16]

2.1.2 RNA Sequencing Technologies and Quantitative Representation of Gene Expression

RNA sequencing (RNAseq) has become the standard technology for transcriptome-wide quantification of gene expression. By converting RNA molecules into complementary DNA (cDNA) fragments and sequencing them using high-throughput platforms, RNAseq enables quantitative measurement of transcript abundance across thousands of genes simultaneously.

Bulk RNAseq and Population-Level Expression Bulk RNAseq measures gene expression averaged across a population of cells. While this approach does not resolve single-cell heterogeneity, it captures robust covariance patterns across genes, which are essential for downstream regulatory network inference.

The typical RNAseq workflow includes RNA extraction, library preparation, sequencing, read alignment to a reference genome, and expression quantification. Alignment tools such as STAR are commonly used to map sequencing reads to genomic coordinates, generating raw count matrices representing the number of reads assigned to each gene.

Figure 2.1: Schematic overview of the RNAseq workflow, including RNA extraction, library preparation, sequencing, alignment, and expression quantification.

Expression Quantification Metrics Gene expression levels can be represented in different formats. The most common include raw counts, Fragments Per Kilobase per Million mapped reads (FPKM), and Transcripts Per Million (TPM).

FPKM is defined as:

$$\text{FPKM} = \frac{\text{Fragments}}{\text{Gene length (kb)} \times \text{Total mapped reads (millions)}} \tag{2.1}$$

While FPKM normalizes for gene length and sequencing depth, it is not optimal for between-sample comparisons. TPM addresses some of these limitations by ensuring that transcript abundances sum to a constant value within each sample.

For regulatory network inference, raw counts followed by appropriate normalization (e.g., library size correction and variance stabilization) are often

preferred, as covariance structures between genes are sensitive to preprocessing choices.

Figure 2.2: Conceptual illustration of how different normalization strategies influence gene expression distributions and downstream covariance structure.

Transcriptomic Covariance and Network Reconstruction A central property of transcriptomic data is the presence of structured statistical dependencies among genes. Co-expressed genes often participate in shared biological pathways or are co-regulated by common transcription factors. These dependencies can be represented through correlation matrices or more sophisticated probabilistic models.

Formally, given an expression matrix:

$$X \in \mathbb{R}^{n \times p} \tag{2.2}$$

where n denotes samples and p genes, regulatory network inference seeks to estimate an adjacency matrix:

$$A \in \mathbb{R}^{p \times p} \tag{2.3}$$

such that A_{ij} captures the regulatory influence of gene i on gene j.

This transformation from expression matrix to interaction network constitutes a shift from feature-based modelling to structure-aware modelling.

Single-Cell RNAseq and Heterogeneity Single-cell RNA sequencing (scRNAseq) extends transcriptomic profiling to individual cells, enabling direct measurement of cellular heterogeneity [**?**]. Although scRNAseq data exhibit higher technical noise and dropout effects compared to bulk RNAseq, they highlight the intrinsic variability underlying population-averaged measurements.

Figure 2.3: Conceptual comparison between bulk RNAseq (averaged signal across cells) and single-cell RNAseq (cell-resolved heterogeneity).

In this thesis, bulk RNAseq data are employed to reconstruct GNNs at the patient level. The statistical dependencies extracted from these measurements serve as the foundation for inferring regulatory interactions and constructing graph-structured representations suitable for downstream GNN modelling.

2.1.3 RNAseq Technologies and Quantitative Representation of Gene Expression

RNA sequencing (RNAseq) provides high-throughput quantification of transcript abundance across thousands of genes simultaneously. In bulk RNAseq experiments, sequencing reads are aligned to a reference genome, and the number of reads mapped to each gene is summarized to produce a gene expression matrix.

Expression Matrix Formalization Formally, RNAseq data can be represented as:

$$X \in \mathbb{R}^{n \times p} \tag{2.4}$$

where n denotes the number of samples (patients) and p denotes the number of genes. Each entry X_{ij} corresponds to the quantified expression level of gene j in sample i.

This matrix representation constitutes the standard input for downstream statistical and ML models.

Quantification Metrics Gene expression can be represented using different normalization schemes, including:

- Raw read counts
- Fragments Per Kilobase per Million mapped reads (FPKM)
- Transcripts Per Million (TPM)

FPKM is defined as:

$$\text{FPKM}_{ij} = \frac{\text{Fragments}_{ij}}{\text{GeneLength}_j \times \text{TotalMappedReads}_i} \times 10^9 \tag{2.5}$$

TPM further rescales expression values so that the total expression within each sample sums to a constant value.

While these transformations adjust for sequencing depth and gene length, they alter the variance structure of the data. Since regulatory network inference depends on covariance relationships among genes, preprocessing choices directly influence the topology of inferred networks.

Covariance Structure and Network Inference A key property of transcriptomic data is the presence of structured statistical dependencies across genes. Let Σ denote the sample covariance matrix:

$$\Sigma = \frac{1}{n-1}(X - \bar{X})^\top (X - \bar{X}) \tag{2.6}$$

where $\bar{X}$ represents the mean-centered matrix.
GNN inference methods attempt to estimate an adjacency matrix:

$$A \in \mathbb{R}^{p \times p} \tag{2.7}$$

where A_{jk} represents the regulatory influence between gene j and gene k. In this context, transcriptomic measurements serve as indirect observations of latent regulatory interactions.

Figure 2.4: Conceptual transformation from an expression matrix (samples $\times$ genes) to a GNN represented as a weighted adjacency matrix.

2.2 Cancer as a Regulatory Network Disease

2.2.1 Network Medicine and Systems-Level Disease Modelling

From Single-Gene Paradigm to Network Perspective Traditional biomedical research has historically adopted a reductionist paradigm, focusing on individual genes or isolated molecular alterations as primary drivers of disease. However, complex diseases such as cancer rarely result from single genetic events. Instead, they emerge from perturbations affecting interconnected molecular systems.

Network medicine provides a conceptual and mathematical framework to interpret disease as a disruption of biological networks rather than isolated molecular defects. In this view, genes, proteins, and regulatory elements are represented as nodes in a graph, while their physical, functional, or regulatory interactions are represented as edges. Disease phenotypes arise from alterations in specific regions of these networks, often referred to as *disease modules* [10].

As described in the network medicine paradigm [10], genes associated with the same disease tend to cluster within localized regions of the interactome, suggesting that pathogenesis reflects coordinated dysfunction of interconnected components rather than independent gene-level failures.

Cancer as a Network Perturbation Acute leukemias exemplify the systems-level nature of oncogenic processes. Large-scale genomic studies, such as The Cancer Genome Atlas (TCGA) analysis of adult de novo acute myeloid leukemia (AML), demonstrated that individual patients harbor combinations of mutations across multiple functional categories, including transcription factors, signaling pathways, DNA methylation regulators, chromatin modifiers, cohesin complex genes, and spliceosome components [**?**].

Importantly, nearly all AML samples contain at least one mutation in genes belonging to a limited number of recurrent functional categories, reinforcing the idea that leukemia is driven by coordinated disruption of regulatory programs rather than by isolated alterations.

These findings support a network-based interpretation of leukemia, where oncogenic transformation reflects the rewiring of transcriptional and regulatory interactions at a systems level.

Transcriptomics and Disease Modules High-throughput transcriptomic profiling has further reinforced the network view of acute leukemia. Whole-transcriptome sequencing (RNAseq) enables not only the identification of gene expression signatures but also the detection of gene fusions and pathway-level alterations [54].

Rather than focusing solely on differential expression of individual genes, transcriptomic data reveal coordinated shifts in gene expression programs, reflecting altered regulatory states. These programs can be interpreted as network modules corresponding to specific leukemia subtypes or risk categories.

Such observations motivate modelling strategies that explicitly incorporate gene-gene relationships and regulatory interactions, rather than treating genes as independent features.

Graph-Based Modelling of Multi-Omics Systems The integration of multi-omics data (genomics, transcriptomics, epigenomics, proteomics) introduces additional complexity, as each layer represents a distinct but interconnected component of the molecular system.

Graph-based approaches, and particularly GNNs (GNNs), have emerged as powerful tools to model such structured biological data. By representing omics entities as nodes and biological relationships as edges, GNNs allow for message passing mechanisms that capture higher-order dependencies and non-linear interactions [105].

In cancer research, GNN-based architectures have been applied to tasks including subtype classification, prognosis prediction, and biomarker discov-

ery, demonstrating improved performance compared to models that ignore network structure.

These developments highlight a methodological convergence between network medicine and graph-based Deep Learning (DL).

Relevance to Patient-Specific Regulatory Networks While global interactomes provide a static representation of molecular relationships, disease manifestation is patient-specific. Inter-patient heterogeneity implies that each individual may exhibit a distinct rewiring of regulatory interactions.

This motivates the reconstruction of patient-specific gene regulatory networks (GRNs), where edge weights reflect sample-specific regulatory strengths. Such representations move beyond bulk-level association and enable modelling of inter-patient variability in regulatory topology.

Within this framework, diseases such as leukemia can be conceptualized as network states embedded in a high-dimensional regulatory space. The transition from normal hematopoiesis to leukemic transformation can thus be interpreted as a shift in network configuration rather than merely a change in gene expression magnitude.

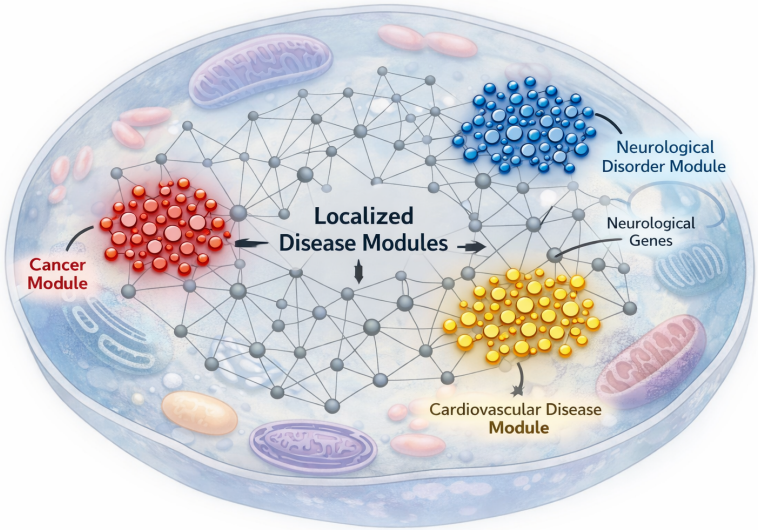


Figure 2.5: Conceptual illustration of disease modules within the human interactome. Disease-associated genes cluster within localized network regions rather than being randomly distributed. Illustration generated with AI assistance and refined by the author.

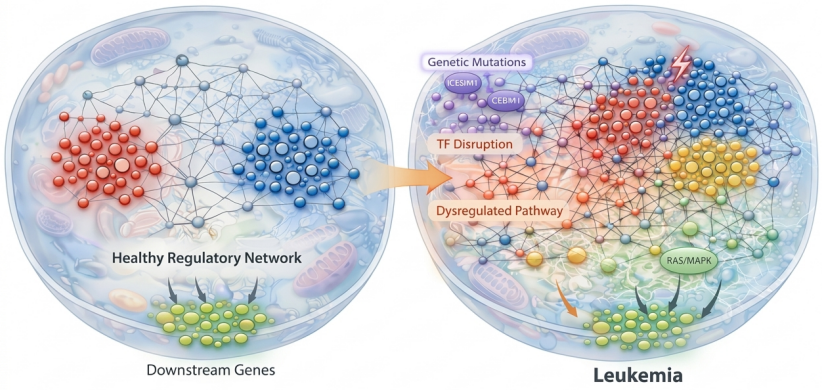


Figure 2.6: Conceptual illustration of regulatory network rewiring in leukemia. In healthy cells (left), transcription factors coordinate structured gene regulatory modules leading to stable downstream gene programs. In leukemia (right), genetic alterations, transcription factor disruption, and signaling pathway dysregulation perturb regulatory interactions, resulting in widespread rewiring of transcriptional programs.

2.2.2 Leukemia: Molecular Subtypes and Regulatory Heterogeneity

Leukemia comprises a heterogeneous group of hematological malignancies originating from transformed hematopoietic progenitor cells. It is broadly classified into acute myeloid leukemia (AML), B-cell acute lymphoblastic leukemia (B-ALL), and T-cell acute lymphoblastic leukemia (T-ALL), based on lineage specification and differentiation stage.

Epidemiological Overview. Leukemia accounts for approximately 2-3% of all malignant tumors worldwide [43]. According to recent global estimates, there were nearly 475,000 new leukemia cases in 2020, corresponding to an age-standardized incidence rate of roughly 5.4 per 100,000 individuals [43].

In Italy, national burden estimates indicate about 15,600 new cases in 2023, with higher incidence rates observed in males than females [69]. Leukemia is the most common malignancy in children, representing approximately 25-30% of all pediatric cancers worldwide [21]. In Italian pediatric populations, the estimated annual incidence is around 30 cases per million children, cor-

responding to roughly 400 new pediatric leukemia diagnoses per year [7]. Among pediatric leukemias, ALL is predominant, accounting for approximately three-quarters of cases, whereas AML comprises about 15-20% of childhood leukemia diagnoses [21].

In adults, the epidemiological landscape differs. AML is the most common acute leukemia in young adults and older individuals, while chronic lymphocytic leukemia (CLL) represents roughly 30% of leukemia cases in adults [8].

Molecular Complexity and Genetic Architecture. Large-scale sequencing studies have revealed that both AML and ALL are characterized by diverse combinations of genetic and epigenetic alterations, including chromosomal rearrangements, gene fusions, point mutations, copy number variations, and aberrant methylation profiles. In the TCGA analysis of 200 adult de novo AML cases, recurrent mutations were identified in transcription factors, DNA methylation regulators, signaling pathways, chromatin modifiers, and spliceosome components, illustrating the combinatorial nature of leukemogenesis [2].

Transcriptomic profiling studies have further enabled refined classification of leukemia subtypes. RNA sequencing not only improves detection of oncogenic fusions but also identifies expression-based subgroups that correlate with clinical risk and therapeutic response [54]. Recent integrative frameworks that combine RNAseq with clinical and cytogenetic data demonstrate improved stratification of previously unresolved cases [86].

Regulatory Heterogeneity and Transcriptional Programs. From a systems biology perspective, leukemia is not adequately explained by isolated genetic lesions. Rather, it reflects altered transcriptional regulatory programs affecting coordinated sets of genes. Dysregulation of key transcription factors (e.g., *RUNX1*, *CEBPA*, *PAX5*, *MYC*) and abnormal activity of epigenetic modifiers (e.g., *DNMT3A*, *TET2*, *IDH1/2*) collectively reshape GNNs (GRNs) and influence cellular phenotype.

Although transcriptomic profiling has enabled molecular subtype stratification, most current classification models treat genes as independent predictors. Supervised ML classifiers based on gene expression matrices, such as ensemble or super learner models, achieve high classification accuracy but do not explicitly model structured regulatory dependencies among genes [79].

Clinical Implications and Need for Network-Aware Modeling. Accurate molecular classification has direct clinical relevance for prognosis, risk

stratification, and therapy selection. Specific genetic events, such as gene fusions or risk-defining mutations, influence treatment choices and targeted therapy eligibility. Furthermore, measurable residual disease (MRD) assessments and transcriptomic risk scores are increasingly incorporated into clinical workflows [86].

Despite these advances, a significant methodological gap persists: the majority of computational frameworks operate on flat gene expression matrices, ignoring the underlying network structure of biological regulation. Considering that leukemia is fundamentally a systems-level disorder involving transcriptional rewiring, modeling patient-specific GNNs offers a promising avenue for capturing inter-patient regulatory heterogeneity and linking genetic alterations to functional regulatory architecture.

2.2.3 Cross-Cancer Disease Association and Shared Regulatory Programs

Cross-cancer disease association analysis aims to identify molecular and regulatory mechanisms that are conserved across distinct tumour types, while preserving the ability to discriminate disease-specific phenotypes. This perspective is consistent with systems biology and network medicine, where cancer is interpreted as a perturbation of interconnected molecular networks and where distinct tumours may share overlapping dysregulated modules (e.g., proliferation, immune evasion, DNA repair) despite arising from different tissues [11].

Large-scale pan-cancer initiatives have demonstrated that tumours can exhibit shared molecular patterns across lineages. In particular, the TCGA Pan-Cancer analysis project was explicitly designed to study commonalities and differences across tumour types by integrating multi-modal molecular profiles [91]. Later Pan-Cancer Atlas efforts further reported that global clustering patterns often reflect cell-of-origin, yet cross-lineage similarities in oncogenic programs and signaling pathways can still be detected across multiple data modalities [41].

Breast Cancer Context and TCGA Resource Availability. Breast cancer is a highly prevalent epithelial malignancy characterized by marked inter-patient molecular heterogeneity, including well-established transcriptomic subtypes. A landmark TCGA study provided a comprehensive molecular portrait of breast tumours, integrating multiple omics layers and supporting subtype stratification and biological interpretation [87]. From a data availability standpoint, breast invasive carcinoma is extensively represented

in the Genomic Data Commons (GDC) portal under the `TCGA-BRCA` project, which includes a large number of RNAseq samples and associated clinical metadata [66].

From Cross-Cancer Similarity to Shared Regulatory Programs. While cross-cancer studies are commonly performed at the feature level (e.g., shared expression signatures), a disease association paradigm can also be formulated at the network level. In this setting, similarities are sought in the configuration of dysregulated regulatory interactions (i.e., GRN modules and hub regulators), rather than in individual gene expression values alone. This is relevant because different tumours may converge to similar phenotypes through distinct genetic events that nevertheless rewire overlapping regulatory circuits.

Recent methodological surveys highlight an increasing trend toward graph-based and network-aware modelling in cancer, including approaches that exploit patient-specific graphs and knowledge-driven priors to capture molecular interactions [105]. Within this framework, testing a unified classifier across leukemia and breast cancer provides a stringent evaluation of whether a regulatory-network-driven representation can capture shared oncogenic programs across tissues while maintaining discriminative performance between normal and tumour states.

Figure 2.7: Schematic of cross-cancer disease association analysis. Distinct tumour types may share dysregulated regulatory modules (e.g., proliferation, immune evasion) despite tissue-specific contexts.

2.3 GNNs and Single-Sample Inference

2.3.1 Formal Representation of GNNs

GNNs (GRNs) are mathematical abstractions designed to represent transcriptional regulation at a systems level. Formally, a GRN can be modeled as a directed, weighted graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}),$$

where nodes $v \in \mathcal{V}$ represent genes or transcription factors (TFs), and directed edges $(i, j) \in \mathcal{E}$ encode regulatory interactions from regulator i to target gene j.

The network can be represented by an adjacency matrix

$$A \in \mathbb{R}^{n \times n}, \tag{2.8}$$

where n denotes the number of genes (or TF-gene pairs), and each entry A_{ij} quantifies the regulatory influence exerted by gene i on gene j. Depending on the inference framework, A_{ij} may encode:

- the existence of a regulatory relationship (binary structure),
- the direction of regulation,
- the type of regulation (activation or inhibition),
- the strength or confidence of the regulatory effect.

As discussed in recent methodological reviews [52], GRN inference can be understood at four progressive levels: (i) detecting whether a regulatory interaction exists, (ii) identifying the regulator and its target, (iii) determining the regulatory sign (activating or inhibitory), and (iv) estimating the magnitude of regulatory strength. Different reconstruction paradigms address different subsets of these levels.

Undirected Association Networks. Correlation-based and mutual information based approaches (e.g., Pearson correlation, ARACNE, CLR) infer undirected or symmetric association networks [16]. In these models, $A_{ij} = A_{ji}$ typically reflects statistical dependency rather than mechanistic causality. While computationally efficient, such networks capture co-expression structure but do not explicitly encode directionality or regulatory logic [**?**, 16]..

Regression and Tree-Based Models. Regression-based and ensemble tree-based methods (e.g., GENIE3, TIGRESS) formulate GRN reconstruction as a feature selection problem, where the expression of a target gene is predicted from candidate regulators [52]. In matrix form, this corresponds to estimating row-wise parameter vectors

$$A_{:j} = \hat{\beta}_j,$$

where $\hat{\beta}_j$ denotes the inferred influence of all potential regulators on gene j. These approaches introduce directionality but often remain limited to statistical associations without modeling explicit regulatory mechanisms.

Dynamical and Boolean Network Models. Ordinary differential equation (ODE)-based models and Boolean networks introduce dynamical structure into GRNs. In ODE formulations, gene expression evolves continuously:

$$\frac{dx_j(t)}{dt} = f_j(x_1(t), \ldots, x_n(t)),$$

where f_j encodes regulatory influences. Boolean network models instead discretize expression states into binary variables and update each node according to logical rules [52]. These approaches allow interpretation of regulatory logic but may require temporal or pseudo-temporal data.

Deep Learning-Based GRN Inference. Recent DL methods extend GRN inference to high-dimensional single-cell RNAseq data, where dropout events and sparsity complicate traditional modeling [16]. Neural architectures aim to capture nonlinear higher-order dependencies and complex regulatory interactions. However, interpretability and biological prior integration remain active research challenges.

Prior-Integrated and Message-Passing Frameworks. A distinct class of methods integrates prior biological knowledge such as TF binding motifs or protein-protein interactions with expression data. These frameworks explicitly encode regulatory directionality and often produce weighted, directed adjacency matrices informed by biological constraints.

Among these, PANDA (Passing Attributes between Networks for Data Assimilation) introduces a message-passing strategy to reconcile motif priors, protein-protein interactions, and gene co-expression into a unified regulatory network. LIONESS (Linear Interpolation to Obtain Network Estimates for Single Samples) extends this formulation by enabling estimation of sample specific adjacency matrices from aggregate network models. These approaches will be described in detail in the following sections.

2.3.2 PANDA: Message-Passing Integration of Regulatory Priors

Passing Attributes between Networks for Data Assimilation (PANDA) is a network reconstruction framework designed to integrate multiple sources of biological information into a unified GRN [33]. Unlike purely data-driven association methods, PANDA incorporates prior knowledge about transcription factor (TF) binding alongside gene expression data and, optionally, protein-protein interaction (PPI) networks.

Input Networks and Notation. PANDA operates on three input matrices:

- A regulatory prior matrix $M \in \mathbb{R}^{n_{TF} \times n_G}$, typically derived from TF binding motif information, where M_{ij} encodes the prior likelihood that TF i regulates gene j.
- A gene co-expression matrix $C \in \mathbb{R}^{n_G \times n_G}$, computed from transcriptomic data (e.g., Pearson correlation).
- A TF-TF interaction matrix $P \in \mathbb{R}^{n_{TF} \times n_{TF}}$, representing protein-protein interactions among TFs.

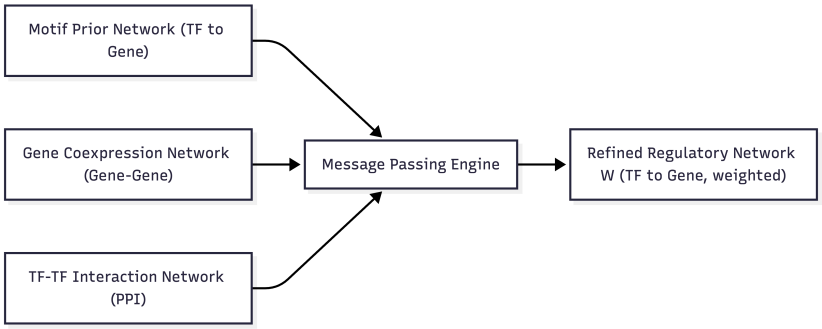


Figure 2.8: Conceptual overview of PANDA. Regulatory priors, gene co-expression, and TF-TF interaction networks are iteratively reconciled through message passing to produce a refined weighted GRN.

The goal is to infer a refined regulatory matrix

$$W \in \mathbb{R}^{n_{TF} \times n_G},$$

where W_{ij} represents the inferred regulatory strength from TF i to gene j after data assimilation.

Message Passing Principle. The core assumption of PANDA is that if two genes are co-expressed, they are more likely to be regulated by similar TFs, and if two TFs interact, they are more likely to regulate similar target genes [33].

PANDA formalizes this intuition through an iterative message-passing procedure that enforces consistency across the three networks. At each iteration, the regulatory matrix W is updated by comparing:

- the similarity between TF regulatory profiles,

- the similarity between gene target profiles,
- the compatibility with the regulatory prior M.

Conceptually, PANDA seeks a regulatory matrix W that maximizes agreement between:

$$\text{TF similarity induced by } W \quad \approx \quad P,$$

$$\text{Gene similarity induced by } W \quad \approx \quad C.$$

The update rules rely on similarity measures (e.g., normalized inner products) computed between rows and columns of W, which are iteratively adjusted until convergence.

Optimization Interpretation. Although PANDA is implemented via iterative message passing, it can be interpreted as solving a constrained optimization problem that balances three objectives:

1. Consistency with motif-based regulatory prior M,
2. Consistency with gene co-expression structure C,
3. Consistency with TF-TF interaction structure P.

The resulting adjacency matrix W is dense and weighted, with edge weights reflecting relative regulatory influence rather than binary relationships.

Properties of PANDA Networks. PANDA-derived GRNs exhibit several distinctive characteristics:

- Directed TF $\rightarrow$ gene structure,
- Continuous edge weights (regulatory strength),
- Integration of heterogeneous biological data sources,
- Population-level estimation (single aggregate network).

Importantly, PANDA produces a global regulatory network estimated from all samples jointly. As such, it captures regulatory processes shared across the population but does not explicitly model inter-sample variability. This limitation motivates the development of sample-specific extensions such as LIONESS, described in the following section.

2.3.3 LIONESS: Sample-Specific Network Estimation

Linear Interpolation to Obtain Network Estimates for Single Samples (LIONESS) is a general framework for deriving sample-specific networks from population-level network inference methods [48]. Its central motivation is that aggregate regulatory networks estimated from a cohort capture interactions shared across the population, but can mask context-specific regulatory variation that is critical for studying inter-patient heterogeneity in complex diseases such as cancer [48].

From Aggregate to Sample-Specific Networks. Let N denote the number of samples in a dataset, and let an aggregate network inference method (e.g., PANDA) produce an edge-weight matrix

$$W^{(\text{all})} \in \mathbb{R}^{n_{TF} \times n_G},$$

estimated using all N samples. Let

$$W^{(\text{all}\setminus q)} \in \mathbb{R}^{n_{TF} \times n_G}$$

denote the aggregate network reconstructed from the same method after removing sample q. LIONESS estimates the sample-specific regulatory network for sample q as a linear interpolation of these two aggregate networks:

$$W^{(q)} = N \cdot W^{(\text{all})} - (N-1) \cdot W^{(\text{all}\setminus q)}. \tag{2.9}$$

Equivalently, writing this edge-wise:

$$w_{ij}^{(q)} = N \left(w_{ij}^{(\text{all})} - w_{ij}^{(\text{all}\setminus q)} \right) + w_{ij}^{(\text{all}\setminus q)}, \tag{2.10}$$

where $w_{ij}^{(q)}$ represents the inferred regulatory influence from TF i to gene j in sample q.

Interpretation and Assumptions. LIONESS relies on the assumption that the aggregate edge weight for a cohort can be expressed as an average contribution of the sample-specific networks generated by the same inference procedure. Under this assumption, the difference between the full-cohort network and the leave-one-out network approximates the contribution of sample q to each edge weight, enabling reconstruction of a complete weighted network per individual [48].

Practical Properties. LIONESS has several practical characteristics that make it suitable for patient-level modeling:

- **Method-agnostic:** it can be applied on top of many aggregate reconstruction methods (e.g., Pearson correlation, mutual information, PANDA) [48].
- **Dense weighted outputs:** it produces a full edge-weight matrix $W^{(q)}$ per sample, enabling downstream analysis of regulatory variability across individuals.
- **Heterogeneity-aware:** it explicitly encodes inter-sample variability in regulatory topology and edge strengths, which is central to studying inter-patient heterogeneity in leukemia.

Relevance for Graph-Based Learning. Within this thesis, LIONESS-derived regulatory matrices provide a principled mechanism to map each patient to an individual GRN. These sample-specific graphs can be used as structured inputs for GNNs, allowing classification and explainability analyses to operate in a regulatory-interaction space rather than on flat gene expression vectors.

Figure 2.9: Leave-one-out principle behind LIONESS. A cohort-level aggregate network is reconstructed on all samples and compared to an aggregate network reconstructed after removing one sample. Their difference is linearly scaled to estimate a sample-specific network.

Figure 2.10: From one cohort to many graphs. LIONESS produces one weighted regulatory network per patient, enabling analysis of regulatory heterogeneity and graph-based learning across individuals.

2.4 ML for Transcriptomic Cancer Classification

ML (ML) has become a central component of computational oncology, particularly for the classification of cancer subtypes using high-dimensional transcriptomic data. In RNAseq-based studies, each sample is typically represented as a vector

$$x \in \mathbb{R}^p,$$

where p denotes the number of genes and $p \gg n$ (number of samples). This high-dimensional, low-sample-size regime is characteristic of biomedical datasets and motivates the use of regularized and ensemble-based learning strategies.

Classical ML approaches operate on tabular representations of gene expression matrices, where each gene is treated as an independent feature. Despite ignoring structured regulatory dependencies, these methods have demonstrated strong predictive performance in many medical classification tasks, including leukemia subtype stratification [31, 70, 79].

2.4.1 Classical ML Approaches

Classical ML models are particularly suitable for biomedical datasets due to:

- Robustness in high-dimensional settings,
- Built-in feature selection mechanisms,
- Interpretability via feature importance measures,
- Strong theoretical foundations.

In transcriptomic cancer classification, two of the most widely adopted models are RFs and SVMs (SVMs) [31, 79, 96].

RF

RF (RF) is an ensemble learning method based on aggregating multiple decision trees [12]. Each tree is trained on a bootstrap sample of the data, and at each split, a random subset of features is considered. The final prediction is obtained by majority voting (classification) or averaging (regression).

Given a training set $\{(x_i, y_i)\}_{i=1}^{n}$, RF constructs T trees:

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^{T} f_t(x),$$

where f_t is the prediction of the t-th tree.

Relevance in Transcriptomics. RF is well-suited to gene expression data because:

- It handles large numbers of features without explicit dimensionality reduction.

- It is robust to noise and correlated predictors.
- It provides feature importance measures, which can be biologically interpreted as candidate biomarkers.

In leukemia studies, RF-based classifiers have achieved high accuracy in subtype prediction using gene expression profiles [79].

Figure 2.11: Schematic representation of RF. Multiple decision trees trained on bootstrapped data are aggregated via majority voting.

SVMs

SVMs (SVMs) are margin-based classifiers that aim to find a hyperplane maximizing separation between classes [22]. Given labeled data (x_i, y_i) with $y_i \in \{-1, 1\}$, the linear SVM solves:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \tag{2.11}$$

subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

Here, w defines the separating hyperplane, b is the bias term, ξ_i are slack variables, and C controls the trade-off between margin width and classification error.

Kernel Trick. SVMs can incorporate nonlinear decision boundaries using kernel functions:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j),$$

where $\phi(\cdot)$ maps data into a higher-dimensional feature space.

Relevance in Cancer Classification. SVMs have been widely applied to microarray and RNAseq datasets due to:

- Strong performance in high-dimensional, low-sample-size settings,
- Theoretical guarantees based on margin maximization,
- Flexibility through kernel methods.

In hematological malignancies, SVM classifiers have demonstrated strong performance in distinguishing leukemia subtypes based on gene expression patterns [31].

Figure 2.12: SVM margin maximization. The optimal hyperplane maximizes the margin between two classes, defined by support vectors.

Limitations of Classical ML in Systems Biology. While RF and SVM models achieve strong predictive performance, they treat gene expression features as independent predictors. Consequently, they do not explicitly model structured biological relationships such as gene regulatory interactions. This limitation motivates graph-based learning approaches that incorporate network topology directly into the classification framework.

2.4.2 Deep Learning on Tabular Transcriptomic Data

DL (DL) extends classical ML by stacking multiple nonlinear transformations, enabling hierarchical feature extraction from high-dimensional data. In transcriptomic cancer classification, each sample is typically represented as a tabular vector

$$x \in \mathbb{R}^p,$$

where p corresponds to the number of genes. Deep neural networks process this vector through successive affine transformations and nonlinear activations, learning latent representations that may capture complex gene interactions.

Although transcriptomic data are inherently structured through regulatory interactions, early DL applications treat gene expression as independent features, similarly to classical ML approaches. In this context, fully connected architectures such as Multi-Layer Perceptrons (MLPs) are commonly employed [31, 96].

Multi-Layer Perceptron (MLP)

An MLP is a feed-forward neural network composed of stacked linear layers interleaved with nonlinear activation functions. Given an input vector $x \in \mathbb{R}^p$, a single hidden layer transformation is defined as:

$$h = \sigma(W_1 x + b_1), \tag{2.12}$$

where $W_1 \in \mathbb{R}^{d \times p}$ is the weight matrix, b_1 is a bias vector, and $\sigma(\cdot)$ is a nonlinear activation function (e.g., ReLU). The output layer then computes:

$$\hat{y} = \text{softmax}(W_2 h + b_2), \tag{2.13}$$

for multi-class classification.

Stacking multiple hidden layers yields deeper architectures capable of learning increasingly abstract representations:

$$x \to h^{(1)} \to h^{(2)} \to \cdots \to \hat{y}.$$

Relevance to Transcriptomics. MLPs are attractive for gene expression data because:

- They can model nonlinear relationships between genes.
- They scale to high-dimensional inputs.
- They are straightforward to implement using modern frameworks (e.g., PyTorch).

However, in purely tabular settings, MLPs do not explicitly encode biological structure and may require strong regularization to avoid overfitting in low-sample-size regimes.

Figure 2.13: Schematic representation of a Multi-Layer Perceptron (MLP). Input gene expression vectors are transformed through stacked linear layers and nonlinear activations.

Feed-Forward Neural Networks (FFN)

The term Feed-Forward Neural Network (FFN) broadly refers to neural architectures where information flows strictly from input to output without recurrent connections. MLPs are a specific instance of FFNs. In transcriptomic classification, FFNs typically consist of:

- Input layer (gene expression vector),
- Multiple fully connected hidden layers,
- Dropout and batch normalization layers for regularization,
- Final classification layer.

Training is performed via backpropagation, minimizing a cross-entropy loss function:

$$\mathcal{L} = -\sum_{i=1}^{n} y_i \log \hat{y}_i. \tag{2.14}$$

Modern implementations rely on automatic differentiation frameworks such as PyTorch, which provide optimized modules for linear layers, activation functions, and loss computation.

Applications in Cancer Transcriptomics. Deep feed-forward networks have been applied to:

- Leukemia subtype prediction,
- Pan-cancer classification,
- Biomarker discovery from RNAseq datasets.

Reviews in computational oncology highlight that while deep models may outperform classical ML in certain settings, performance gains are often modest when biological structure is not explicitly encoded [31, 95]. This observation motivates structured DL approaches, such as GNNs, that integrate prior biological knowledge.

Figure 2.14: Feed-forward training process. Gene expression inputs are propagated forward through the network; gradients are computed via backpropagation to update weights.

Limitations in Systems-Level Modeling. Despite their flexibility, tabular DL models treat genes as independent coordinates in Euclidean space. They do not incorporate explicit regulatory topology, interaction structure, or network constraints. Consequently, they may capture statistical dependencies but not mechanistic regulatory rewiring. This limitation motivates the transition toward graph-based representations of transcriptomic data.

2.5 Graph Neural Networks for Regulatory Network Modeling

2.5.1 Motivation for Graph-Based Learning

GRNs naturally define graph-structured data. In a GRN, transcription factors and genes form nodes, while regulatory interactions define directed,

weighted edges. Unlike tabular transcriptomic representations, where each gene is treated as an independent feature, GRNs encode explicit relational dependencies between molecular entities.

Formally, a regulatory graph can be defined as:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}),$$

where $\mathcal{V}$ is the set of nodes (genes or TFs) and $\mathcal{E}$ is the set of edges. Let $A \in \mathbb{R}^{n \times n}$ denote the adjacency matrix, where A_{ij} represents the regulatory strength from node i to node j. Each node may additionally be associated with a feature vector $x_i \in \mathbb{R}^d$ (e.g., gene expression levels).

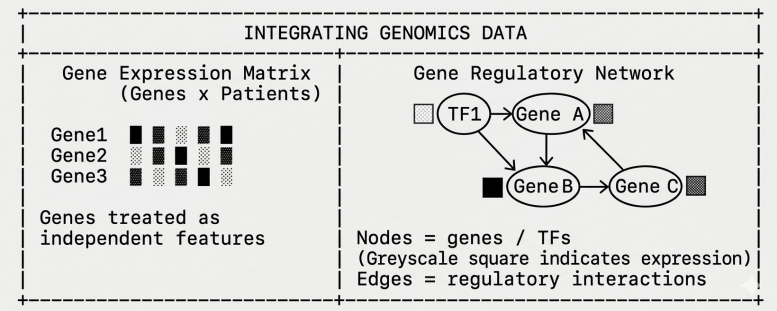


Figure 2.15: Tabular versus graph-based representation of transcriptomic data. Traditional transcriptomic datasets are represented as gene expression matrices (genes Ã patients), where genes are treated as independent features. In contrast, gene regulatory networks model genes and transcription factors as nodes connected by regulatory interactions, explicitly encoding dependencies between molecular entities.

GNNs. Graph neural networks extend DL to non-Euclidean domains by explicitly operating on graphs. Instead of learning transformations of independent feature vectors, GNNs propagate information across edges using iterative neighborhood aggregation.

A general message-passing layer can be expressed as:

$$h_i^{(k+1)} = \phi\left(h_i^{(k)}, \operatorname{AGG}_{j \in \mathcal{N}(i)} \psi(h_i^{(k)}, h_j^{(k)}, A_{ij})\right), \tag{2.15}$$

where:

- $h_i^{(k)}$ is the node representation at layer k,
- $\mathcal{N}(i)$ denotes the neighborhood of node i,

- $\psi(\cdot)$ computes messages along edges,
- AGG aggregates messages (e.g., sum, mean, max),
- $\phi(\cdot)$ updates node embeddings.

Through repeated propagation, nodes integrate multi-hop contextual information, enabling the model to capture higher-order dependencies beyond direct neighbors.

Relevance to GRNs and Cancer Heterogeneity. In regulatory networks:

- Transcription factors influence sets of downstream genes.
- Genes may share regulators, forming regulatory modules.
- Dysregulated hubs can propagate perturbations across the network.

GNNs provide a principled mechanism to model such structured dependencies. When applied to patient-specific GRNs derived via LIONESS, each patient is represented as a graph instance, enabling classification in regulatory-interaction space rather than in flat expression space.

From Message Passing to Graph-Level Prediction. For graph classification tasks, node embeddings are typically aggregated into a graph-level representation:

$$h_{\mathcal{G}} = \text{READOUT}\left(\{h_i^{(K)}\}_{i\in\mathcal{V}}\right), \tag{2.16}$$

where READOUT may correspond to global mean pooling or attention-based pooling. The resulting embedding is then fed into a classifier for disease subtype prediction.

This general message-passing framework encompasses several specific architectures, including Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and GraphSAGE, which differ in how neighborhood aggregation and edge weighting are implemented. These architectures are described in the following sections.

2.5.2 Graph Convolutional Networks (GCN)

Graph Convolutional Networks (GCNs) extend convolutional operations to graph domains by performing neighborhood-based feature aggregation [47]. Unlike classical convolution defined on regular grids, graph convolution operates on irregular graph structures through adjacency-based propagation.

Given an adjacency matrix A and node feature matrix $H^{(k)} \in \mathbb{R}^{n \times d_k}$ at layer k, the GCN update rule is:

$$H^{(k+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(k)}W^{(k)}\right), \tag{2.17}$$

where:

- $\tilde{A} = A + I$ adds self-loops,
- $\tilde{D}$ is the degree matrix of $\tilde{A}$,
- $W^{(k)}$ is a learnable weight matrix,
- $\sigma(\cdot)$ is a nonlinear activation function.

This symmetric normalization ensures numerical stability and avoids scale explosion in deep layers.

Interpretation in GRNs. In regulatory networks, GCN layers allow each gene node to aggregate information from its regulatory neighbors. Through multiple layers, information propagates across multi-hop regulatory paths, enabling the model to capture indirect regulatory effects.

Figure 2.16: Graph Convolutional Network layer. Node embeddings are updated by normalized aggregation of neighboring features followed by linear transformation.

Limitations. GCNs assume uniform weighting of neighbors (after normalization) and may suffer from over-smoothing when many layers are stacked, leading to indistinguishable node embeddings.

2.5.3 Graph Attention Networks (GAT)

Graph Attention Networks introduce an attention mechanism to learn adaptive weights for neighbor contributions [89]. Instead of fixed normalization, GAT computes attention coefficients for each edge.

For node i and its neighbor j, the attention coefficient is:

$$e_{ij} = \text{LeakyReLU}\left(a^\top [Wh_i \,||\, Wh_j]\right), \tag{2.18}$$

where a is a learnable vector and $||$ denotes concatenation. The coefficients are normalized using softmax:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})}. \tag{2.19}$$

The node update becomes:

$$h_i^{(k+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} W h_j\right). \tag{2.20}$$

Relevance for Regulatory Networks. In GRNs, not all regulatory interactions contribute equally. GAT allows the model to assign higher importance to biologically relevant regulatory edges, potentially highlighting influential transcription factors or dysregulated hubs.

Figure 2.17: Graph Attention mechanism. Each neighbor contributes with a learned attention weight, allowing adaptive aggregation.

Advantages.

- Adaptive weighting of neighbors.
- Increased expressivity compared to GCN.
- Potential interpretability through attention coefficients.

2.5.4 GraphSAGE

GraphSAGE (Graph Sample and Aggregate) introduces an inductive framework for learning node embeddings [38]. Instead of relying on the full adjacency matrix, GraphSAGE samples fixed-size neighborhoods and aggregates their features.

The generic update rule is:

$$h_i^{(k+1)} = \sigma\left(W^{(k)} \cdot \text{CONCAT}\left(h_i^{(k)}, \text{AGG}\left(\{h_j^{(k)} : j \in \mathcal{N}(i)\}\right)\right)\right), \tag{2.21}$$

where AGG may correspond to mean, max pooling, or LSTM-based aggregation.

Inductive Learning. A key property of GraphSAGE is its ability to generalize to unseen graphs, as embeddings are computed through learned aggregation functions rather than fixed spectral operations.

Application to Patient-Specific GRNs. In the context of LIONESS-derived regulatory networks, each patient graph may exhibit unique topology and edge weights. GraphSAGE enables learning transferable aggregation rules that generalize across patients, making it suitable for graph-level classification tasks.

Figure 2.18: GraphSAGE sampling and aggregation. A fixed-size neighborhood is sampled and aggregated before updating node embeddings.

Comparison of Architectures.

- GCN: normalized spectral aggregation.
- GAT: attention-weighted aggregation.
- GraphSAGE: sampled inductive aggregation.

These architectures represent complementary strategies for modeling regulatory interactions. Their comparative evaluation in leukemia subtype classification forms part of the methodological investigation presented in later chapters.

2.5.5 Edge-Aware and Signed Graph Models

Regulatory networks differ from generic graphs in two critical aspects: (i) edges carry continuous weights reflecting regulatory strength, and (ii) interactions may be signed, representing activation or inhibition. Standard message-passing architectures such as GCN and GraphSAGE primarily focus on structural adjacency and often treat edges as unweighted or uniformly normalized. Edge-aware and signed graph models extend this framework by incorporating richer relational information.

Graph Isomorphism Network (GIN)

The Graph Isomorphism Network (GIN) was introduced to achieve maximal discriminative power among message-passing architectures [92]. Unlike GCN,

which applies normalized averaging, GIN employs a sum-based aggregation followed by a multilayer perceptron (MLP):

$$h_i^{(k+1)} = \mathrm{MLP}^{(k)}\left((1+\epsilon)h_i^{(k)} + \sum_{j\in\mathcal{N}(i)} h_j^{(k)}\right), \tag{2.22}$$

where ϵ is either fixed or learnable.

Advantage for GRNs. The sum aggregator preserves structural distinctions that averaging operations may obscure. In regulatory networks, where small differences in connectivity patterns may correspond to biologically meaningful regulatory rewiring, the higher expressive power of GIN can improve graph-level discrimination between patient subtypes.

Graph Isomorphism Network with Edge Features (GINE)

GINE extends GIN by explicitly incorporating edge features into the aggregation process [42]. The update rule becomes:

$$h_i^{(k+1)} = \mathrm{MLP}^{(k)}\left((1+\epsilon)h_i^{(k)} + \sum_{j\in\mathcal{N}(i)} \phi(h_j^{(k)}, e_{ij})\right), \tag{2.23}$$

where e_{ij} represents edge attributes and ϕ is typically a learnable transformation combining node and edge information.

Relevance for Weighted GRNs. LIONESS-derived regulatory networks produce continuous edge weights reflecting regulatory strength. GINE allows these weights to directly influence message passing, enabling the model to distinguish strong regulatory interactions from weak ones. This is particularly important in biological networks, where edge magnitude carries mechanistic significance.

GRNFormer and Graph Transformer Models

Graph Transformer architectures adapt self-attention mechanisms to graph domains by allowing nodes to attend to each other based on learned relational scores. In the context of GNNs, GRNFormer-type models leverage transformer-style attention to model long-range dependencies and global regulatory context [30].

A generic graph transformer update can be expressed as:

$$\alpha_{ij} = \text{softmax}_j \left(\frac{(W_Q h_i)^\top (W_K h_j)}{\sqrt{d}} + b_{ij} \right), \tag{2.24}$$

$$h_i^{(k+1)} = \sum_{j \in \mathcal{V}} \alpha_{ij} W_V h_j, \tag{2.25}$$

where b_{ij} may encode structural bias derived from graph topology.

Advantages in Regulatory Modeling. Graph transformer models offer several potential benefits:

- Modeling long-range regulatory dependencies beyond local neighborhoods,
- Flexible incorporation of edge weights and structural encodings,
- Enhanced expressivity in heterogeneous biological graphs.

In GRNs, transcription factors may influence distant genes through multi-step regulatory cascades. Transformer-based models can capture such global dependencies more effectively than strictly local message-passing architectures.

Comparative Perspective. The investigated architectures differ in how they treat neighborhood aggregation:

- GIN emphasizes structural expressivity.
- GINE incorporates edge attributes explicitly.
- Graph Transformer models capture global dependencies.

These variants provide complementary modeling capabilities for regulatory networks reconstructed via PANDA and LIONESS, enabling systematic evaluation of architectural suitability for leukemia subtype classification.

2.6 Explainability in Graph-Based Oncology

Explainability is Essential in GRN-Based Models. Unlike image or text domains, regulatory graphs represent real biological interactions. Therefore, explanations derived from GNN models can map directly to:

- Individual genes,

- Regulatory hubs,
- Interaction pathways,
- Patient-specific regulatory perturbations.

Explainability thus serves both scientific discovery and clinical interpretability.

2.6.1 Saliency Maps and Integrated Gradients

Gradient-based attribution methods compute the sensitivity of model output with respect to input features. For a prediction $\hat{y}$ and input feature x_i, saliency is defined as:

$$S_i = \left| \frac{\partial \hat{y}}{\partial x_i} \right|. \tag{2.26}$$

In transcriptomic settings, this corresponds to identifying genes whose expression changes most strongly affect classification output.

Integrated Gradients (IG) improve upon simple gradients by accumulating contributions along a path from a baseline input x' to the actual input x:

$$IG_i(x) = (x_i - x_i') \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha. \tag{2.27}$$

Outputs in Regulatory GNNs. When applied to GNN-based patient graphs:

- Saliency can identify top influential genes per patient.
- Aggregating across patients enables identification of class-specific gene signatures.
- In patient-to-patient similarity networks, it can highlight characteristic individuals per subtype.

Figure 2.19: Gradient-based attribution. Feature importance is estimated via sensitivity of model output to input perturbations.

2.6.2 SHAP and Model-Agnostic Feature Attribution

SHAP (SHapley Additive exPlanations) provides feature attribution based on cooperative game theory [59]. It assigns each feature a contribution value reflecting its marginal impact on prediction.

For a model $f(x)$, SHAP values ϕ_i satisfy:

$$f(x) = \phi_0 + \sum_{i=1}^{p} \phi_i. \tag{2.28}$$

SHAP can be applied in model-agnostic settings or via gradient-based approximations such as those implemented in Captum.

Application in Transcriptomic Classification. In GRN-based classification:

- SHAP identifies globally important genes.
- Class-level aggregation reveals subtype-discriminative biomarkers.
- Captum-based SHAP-like methods allow integration with PyTorch GNN models.

Compared to simple gradients, SHAP provides more stable and theoretically grounded attribution estimates.

Figure 2.20: SHAP attribution. Each gene contributes additively to the final prediction according to Shapley values.

2.6.3 GNNExplainer

GNNExplainer is a model-specific explainability method designed for GNNs [94]. It identifies a compact subgraph and subset of node features that are most influential for a specific prediction.

The method optimizes a mask over edges and features:

$$\max_{M_E, M_X} \; I(Y; G \odot M_E, X \odot M_X), \tag{2.29}$$

where M_E and M_X are learnable masks for edges and node features, respectively.

Interpretation in GRNs. In regulatory networks, GNNExplainer can:

- Identify critical regulatory pathways.
- Highlight transcription factor-target modules driving subtype prediction.
- Reveal patient-specific dysregulated subgraphs.

Unlike feature-level attribution methods, GNNExplainer provides structural explanations, aligning naturally with biological pathway analysis.

Figure 2.21: GNNExplainer concept. A subgraph and feature mask are optimized to explain model predictions.

Comparative Perspective.

- Saliency / IG: feature-level importance (top genes).
- SHAP: additive global and local feature attribution (top genes).
- GNNExplainer: structural subgraph identification (pathways).

Together, these methods provide complementary interpretability perspectives for graph-based oncology models.

2.7 Conceptual Synthesis and Research Gap

RNA expression profiles provide quantitative measurements of gene activity, but they also implicitly encode latent regulatory programs governing cellular behavior. Rather than acting independently, genes operate within structured transcriptional networks coordinated by transcription factors and regulatory modules. From a systems-level perspective, cancer is not merely a consequence of isolated gene alterations but a manifestation of network-level dysregulation.

In hematological malignancies, leukemia subtypes are characterized by distinct transcriptional programs and regulatory rewiring. While transcriptomic profiling has enabled increasingly refined molecular stratification, most predictive models treat gene expression features as independent variables,

thereby neglecting the structured dependencies intrinsic to regulatory biology. Early studies on leukemia classification primarily relied on gene expression profiles derived from microarray technologies, where classical ML models were used to distinguish between acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) [35]. These approaches typically treated genes as independent variables and therefore did not capture the complex regulatory interactions underlying transcriptional programs.

More recent work has explored alternative data modalities such as flow cytometry combined with DL models for automated leukemia subtype identification [19]. While these approaches have shown promising predictive performance, they focus on cellular phenotype measurements rather than transcriptomic regulatory mechanisms.

Artificial Intelligence (AI) frameworks for leukemia diagnostics have also been proposed to integrate genomic and clinical data in decision-support systems [86]. These systems demonstrate the potential of AI-assisted diagnostics but rarely incorporate explicit regulatory network modelling.

GRNs provide a natural formalism to model intrinsic regulatory mechanisms. PANDA enables integration of biological priors with gene expression data to estimate population-level regulatory networks, while LIONESS extends this framework to derive patient-specific GRNs. These sample-specific networks offer structured graph representations capable of encoding inter-patient regulatory heterogeneity.

Graph-based approaches have gained increasing attention in recent years. In particular, GNNs have been applied to model cell-cell communication or tumor microenvironment interactions using single-cell RNAseq data [44]. However, these studies typically construct graphs from similarity metrics or ligand-receptor interactions rather than transcriptional regulatory networks.

More broadly, recent surveys on multi-omics cancer analysis highlight the growing role of GNN architectures in modelling complex biological systems [105]. These works emphasize the importance of integrating structured biological priors to improve interpretability and predictive performance.

GNNs, through message passing and hierarchical representation learning, advanced architectures - including GCN, GAT, GraphSAGE, GIN, and edge-aware variants - can exploit relational information that is inaccessible to classical tabular models. When combined with explainability methods, these models offer the possibility of identifying biologically meaningful genes, regulatory hubs, and dysregulated pathways.

Research Gap. Despite advances in transcriptomic classification and graph-based DL, three major gaps remain:

1. **Limited integration of patient-specific regulatory networks into predictive modeling.** Most cancer classification studies rely on flat expression matrices, while few explicitly incorporate sample-specific GRNs as structured inputs.

2. **Insufficient modeling of inter-patient regulatory heterogeneity.** Aggregate network representations obscure individual regulatory rewiring, which is central to leukemia subtype diversity.

3. **Lack of systematically evaluated, explainable graph-based pipelines in cross-cancer settings.** Although GNNs have been applied in oncology, comprehensive frameworks integrating GRN reconstruction, advanced GNN architectures, and interpretability analysis remain limited, particularly in disease-association paradigms spanning multiple tumor types.

Positioning of This Thesis. This thesis investigates whether integrating LIONESS-derived patient-specific GNNs with advanced GNNs improves leukemia subtype classification while preserving biological interpretability. Furthermore, it evaluates the generalizability of this framework in a cross-cancer disease association setting, testing whether shared oncogenic regulatory programs can be captured across distinct tumor types.

By bridging regulatory network reconstruction, graph-based DL, and explainable modeling, this work aims to advance structured transcriptomic analysis beyond feature-level prediction toward mechanistically informed, patient-aware computational oncology.

Chapter 3

Dataset and Pre-processing

3.1 Data Sources and Cohort Description

3.1.1 Public Repositories and Data Access

The identification of an appropriate transcriptomic cohort is a fundamental prerequisite for constructing reproducible GNNs and graph-based predictive models. The evaluation criteria included: (i) availability of harmonised RNAseq data, (ii) sample size, (iii) presence of detailed clinical annotations, (iv) possibility of multi-omics integration, (v) standardisation of preprocessing pipelines, and (vi) open access availability.

Several major biological data repositories were initially investigated in order to evaluate their suitability for patient-specific regulatory modelling. Among them, protein-protein interaction (PPI) and pathway databases such as STRING [6], BioGRID [4], and KEGG [5], as well as large-scale transcriptomic resources such as the Genotype-Tissue Expression (GTEx) project [3], were examined.

STRING and BioGRID provide curated protein-protein interaction networks, which are valuable for constructing prior biological graphs. KEGG offers pathway-level knowledge that can support functional interpretation and regulatory inference. GTEx provides large-scale RNAseq profiles across healthy tissues and is particularly relevant for studying baseline transcriptional variability. However, these resources lack the combination of tumour-specific transcriptomic data, harmonised multi-omics layers, and detailed patient-level clinical annotations required for modelling inter-patient heterogeneity in cancer.

Following comparative assessment, the **Genomic Data Commons (GDC)** portal was selected as the primary data source. The GDC is an open-access data infrastructure developed by the National Cancer Institute (NCI)

that harmonises large-scale cancer genomic projects, including **The Cancer Genome Atlas (TCGA)** and **Therapeutically Applicable Research to Generate Effective Treatments (TARGET)**. Its harmonised processing pipelines ensure cross-project comparability and minimise technical variability [36].

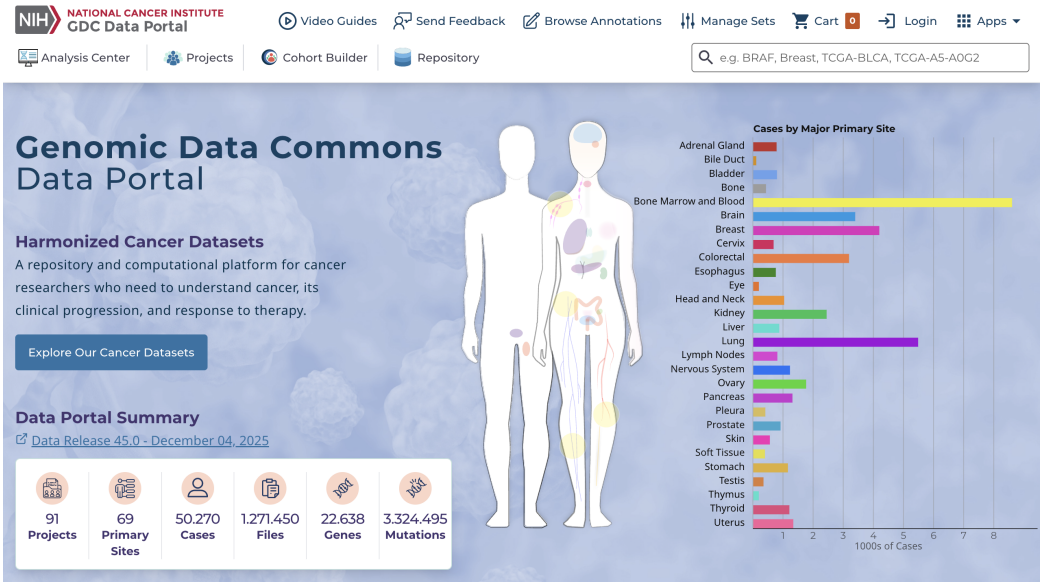


Figure 3.1: Overview of the Genomic Data Commons (GDC) portal interface enabling project exploration, filtering by molecular data type, and download of harmonised datasets.

The GDC portal [65] allows structured exploration of cancer projects, filtering by molecular data type (RNAseq, mutations, methylation, copy number variation), inspection of clinical metadata, and retrieval of both raw and processed files. Importantly, all datasets are processed using standardised pipelines, thereby reducing batch-related artefacts introduced by heterogeneous preprocessing procedures.

The **Cancer Genome Atlas (TCGA)** is a landmark initiative aimed at molecularly profiling more than 11,000 tumours across 33 cancer types. TCGA provides multi-omics datasets including RNAseq, DNA methylation (beta-values), somatic mutations (MAF files), copy number variation (CNV), microRNA sequencing, and proteomics, alongside detailed clinical annotations such as survival, tumour stage, and treatment response. The structured TCGA barcode (e.g., `TCGA-AB-1234-01A-01R`) enables consistent linkage between molecular files and patient-level metadata.

Example: TCGA-AB-1234-01A-01R

Code Segment	Meaning
TCGA	Project
AB-1234	Patient
01	Sample type (tumor/normal)
A-01R	Technical details

Figure 3.2: Structure of the TCGA barcode used to map molecular files to clinical annotations.

TCGA has substantially contributed to the molecular characterisation of acute myeloid leukaemia, identifying recurrent driver mutations, epigenetic alterations, and transcriptional rewiring patterns [2]. Such large-scale profiling efforts provide a robust foundation for regulatory network reconstruction.

Project	Disease Type	Program	Cases	RNAseq
TARGET-AML	Myeloid Leukemias	TARGET	**2492**	✓
TARGET-ALL-P2	Lymphoid Leukemias	TARGET	**1587**	✓
TARGET-NBL	Neuroblastoma	TARGET	1132	✓
TCGA-BRCA	Breast Invasive Carcinoma	TCGA	**1098**	✓
TCGA-GBM	Glioblastoma	TCGA	617	✓
TCGA-OV	Ovarian Carcinoma	TCGA	608	✓
TCGA-LUAD	Lung Adenocarcinoma	TCGA	585	✓
TCGA-UCEC	Endometrial Carcinoma	TCGA	560	✓
TCGA-KIRC	Kidney Renal Clear Cell Carcinoma	TCGA	537	✓
TCGA-HNSC	Head and Neck Squamous Carcinoma	TCGA	528	✓
TCGA-LGG	Lower Grade Glioma	TCGA	516	✓
TCGA-THCA	Thyroid Carcinoma	TCGA	507	✓
TCGA-LUSC	Lung Squamous Cell Carcinoma	TCGA	504	✓
TCGA-PRAD	Prostate Adenocarcinoma	TCGA	500	✓
TCGA-SKCM	Skin Cutaneous Melanoma	TCGA	470	✓
TCGA-COAD	Colon Adenocarcinoma	TCGA	461	✓
TCGA-STAD	Stomach Adenocarcinoma	TCGA	443	✓
TCGA-BLCA	Bladder Carcinoma	TCGA	412	✓

Table 3.1: Major GDC projects initially screened (Projects $>$ 400 cases). Highlighted rows indicate the cohorts selected for this study: TARGET-AML and TARGET-ALL-P2 (primary leukaemia task) and TCGA-BRCA (cross-cancer disease association analysis).

Data acquisition was performed through three complementary modalities:

- the GDC graphical interface for project exploration and manifest generation,
- the GDC REST API for reproducible programmatic queries, and
- the GDC Command Line Interface (CLI) for bulk download of RNAseq files via manifest JSON files.

Harmonised gene-level RNAseq quantifications derived from the HTSeq pipeline were selected. Raw gene-level counts were retained to allow controlled downstream normalisation and consistent integration into the GNN reconstruction framework.

Samples were included based on the availability of RNAseq quantification, corresponding clinical annotations, and primary tumour designation. Duplicated aliquots and technical replicates were removed to avoid redundancy. This selection strategy ensures methodological consistency and provides a robust foundation for constructing patient-specific regulatory graphs.

3.1.2 Leukemia Cohort (Primary Task)

Cohort definition and rationale. The primary task in this thesis focuses on acute leukaemia, leveraging harmonised RNAseq profiles and associated clinical metadata from the **TARGET** programme accessed through the **Genomic Data Commons (GDC)**. TARGET provides large paediatric cohorts and multi-omic profiling, offering a suitable substrate for modelling inter-patient heterogeneity via patient-specific GNNs. In this work, the leukaemia cohort was built by considering two TARGET projects: **TARGET-AML** (acute myeloid leukaemia) and **TARGET-ALL** (acute lymphoblastic leukaemia), both of which include RNAseq and complementary molecular assays. Project-level cohort sizes reported by the GDC portal are: **TARGET-AML: 2,492 cases** and **TARGET-ALL-P2: 1,587 cases** (initial screening, prior to task-specific inclusion filters).Table 3.1. [36, 63, 64]

Clinical context: AML and ALL. Acute myeloid leukaemia (AML) and acute lymphoblastic leukaemia (ALL) represent biologically distinct haematological malignancies characterised by disrupted differentiation and aberrant transcriptional programmes. Modern transcriptomic profiling has become central to improved diagnosis, molecular stratification, and precision medicine in acute leukaemia, supporting both subtype classification and downstream biological interpretation. [54] In AML specifically, large-scale genomics has identified recurrent alterations across signalling pathways, transcriptional regulators, and epigenetic modifiers, consistent with a network-level dysregulation view of the disease. [2]

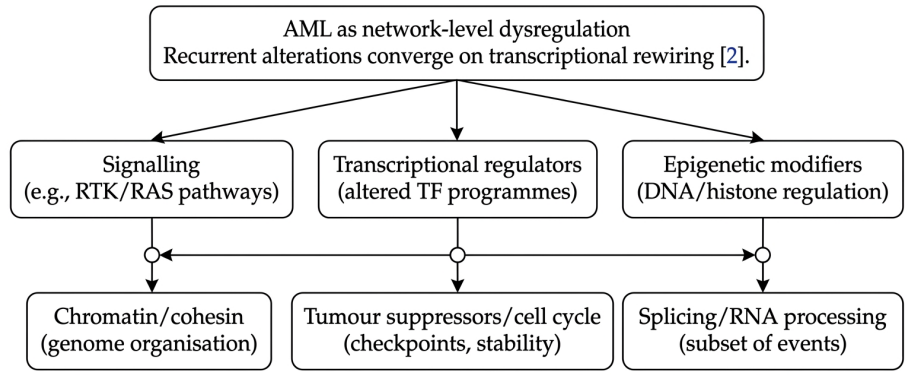


Figure 3.3: Schematic of AML molecular landscape. Conceptual functional groupings of recurrent alterations motivate network-level representations and regulatory-network-based modelling. [2]

Metadata availability in GDC/TARGET. A key practical advantage of using TARGET through the GDC portal is the breadth of patient-level metadata that can be leveraged for cohort refinement and downstream stratified analyses. Available fields include, among others: *tissue type* (tumour vs normal), *specimen type* (e.g., bone marrow, peripheral blood), *primary diagnosis* descriptors (AML/ALL variants), *vital status*, demographic variables (sex, race, ethnicity), and treatment-related fields (treatment outcome, therapeutic agents, treatment type). These metadata, visible in Table. 3.2 were used to define inclusion filters and to support clinically meaningful label construction.

Unique-case selection and prevention of leakage. For graph classification, avoiding multiple samples from the same patient is essential to prevent information leakage across training and evaluation folds. Following the cohort refinement strategy illustrated in the dataset curation diagram (Fig. 3.4), the selection was restricted to **one sample per patient** after applying filters on tissue type and diagnosis, and after removing duplicates. The resulting working cohort used for subsequent modelling contained **2,374 unique samples** (Fig. 3.4). This strategy preserves patient independence and supports robust cross-validation.

Table 3.2: Example metadata annotations available in the GDC portal for leukemia samples.

Metadata category	Example annotations
Tissue type	normal; tumor
Specimen type	bone marrow; peripheral blood; fibroblasts (BM); derived cell line; solid tissue; buccal cells; unknown
Primary diagnosis	acute lymphocytic leukemia (ALL); acute myeloid leukemia (AML); acute monocytic leukemia; acute megakaryoblastic leukemia; acute myelomonocytic leukemia
Gender	female; male; not reported; unknown
Therapeutic agents	gemtuzumab ozogamicin; hydroxyurea; tretinoin
Treatment type	pharmaceutical therapy; stem cell transplantation; steroid therapy

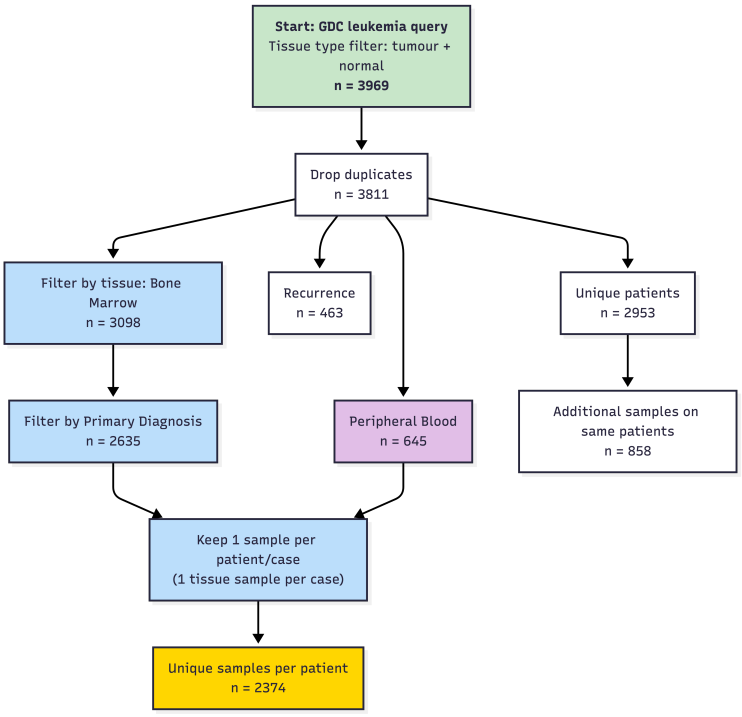


Figure 3.4: Steps executed to select unique patients and support cross-validation.

Class distribution and imbalance analysis. After filtering and retaining one sample per patient, the final working cohort consisted of **2,374 unique cases**. Two supervised learning configurations were derived from this cohort: (i) a binary classification setting (Tumour vs Normal), and (ii) a three-class subtype setting.

Setting	Class	Proportion (%)
Binary (Tumour vs Normal)	Tumour	83.0
	Normal	17.0
Multiclass (Subtype)	AML (Major subtype)	69.3
	ALL	16.0
	Normal	14.7

Table 3.3: Class distribution in the curated leukaemia cohort (n = 2,374).

The binary configuration exhibits a moderate class imbalance ratio ($IR = majority_{class}/minority_{class}$) of 5.2. In the multiclass setting, the largest subtype accounts for 69.3% of samples, while the remaining classes represent 16.0% and 14.7%, respectively.

The distributions differ between the two configurations because the binary setting aggregates all tumour entities into a single class. In addition to Acute Myeloid Leukaemia (AML) and Acute Lymphoblastic Leukaemia (ALL), the tumour category includes low-frequency entities such as *Induction Failure AML (AML-IF)* and other rare diagnostic annotations present in the TARGET metadata. These rare subclasses contribute marginally to the total sample size but affect the tumour proportion in the binary formulation. This imbalance motivates the adoption of stratified cross-validation and class-weighted loss functions, as detailed in Section 3.5.

3.1.3 Cross-Cancer Cohort (Disease Association Paradigm)

Rationale and biological motivation. To evaluate the generalisability of the proposed regulatory-network-based modelling framework beyond haematological malignancies, a second independent tumour cohort was incorporated. Breast Invasive Carcinoma (TCGA-BRCA) was selected from the **The Cancer Genome Atlas (TCGA)** programme within the GDC infrastructure [36, 66].

Breast cancer represents a solid tumour with a distinct tissue origin (Table 3.1), mutational landscape, and transcriptional architecture compared to

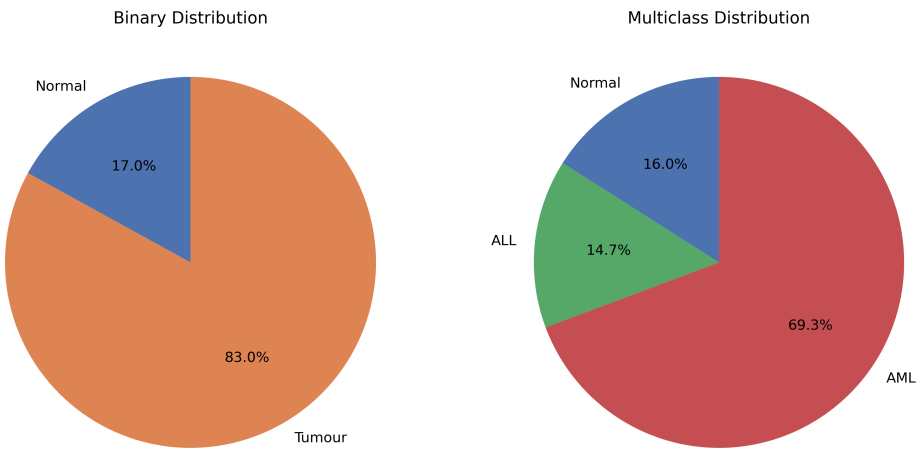


Figure 3.5: Class distribution of the leukemia dataset. Left: binary classification (normal vs tumour). Right: multiclass classification (normal, ALL, AML).

acute leukaemia. The inclusion of TCGA-BRCA enables a **cross-tissue disease association setting**, allowing assessment of whether patient-specific regulatory representations capture shared oncogenic regulatory patterns across biologically heterogeneous tumour types.

This configuration directly operationalises the *disease association modelling* concept introduced in Chapter 2, in which graph-based transcriptomic representations are evaluated across distinct disease contexts rather than within a single tumour lineage.

Cohort size and composition. After applying harmonised RNAseq filters and retaining one sample per patient, the TCGA-BRCA cohort consisted of:

- **1,157 unique samples**
- **1,051 tumour samples**
- **106 normal samples**

This corresponds to a tumour proportion of 90.8% and a normal proportion of 9.2%, yielding a tumour-to-normal imbalance ratio of approximately 9.9:1.

Multi-tumour cohort construction. For the disease association experiment, the binary tumour/normal cohorts from:

- TARGET leukaemia (Section 3.1.2)
- TCGA-BRCA

were combined into a unified multi-tumour dataset.
The resulting dataset integrates transcriptomic profiles from:

- Haematopoietic malignancies (AML, ALL, rare subtypes)
- Solid epithelial tumours (breast carcinoma)

All samples were processed through the same preprocessing and graph construction pipeline to ensure methodological consistency.

Binary classification task (cross-cancer setting). The cross-cancer task is formulated as:

Tumour vs **Normal**

where:

- Tumour includes AML, ALL, rare leukaemia entities (e.g., AML-IF), and breast carcinoma.
- Normal includes matched or control samples from both bone marrow / haematological and breast cohorts.

This formulation evaluates whether GNNs trained on regulatory-network-derived graphs can discriminate malignant from non-malignant states across tissue types.

Combined binary dataset summary. The leukaemia and breast cancer cohorts exhibit distinct class imbalance structures and different tumour-to-normal ratios. Furthermore, their biological heterogeneity and tissue-specific transcriptional variability differ substantially.

To mitigate potential bias induced by unequal cohort sizes and distributional asymmetry, next section introduces controlled sample equalisation strategies based on centrality- or diversity-driven selection criteria prior to model training.

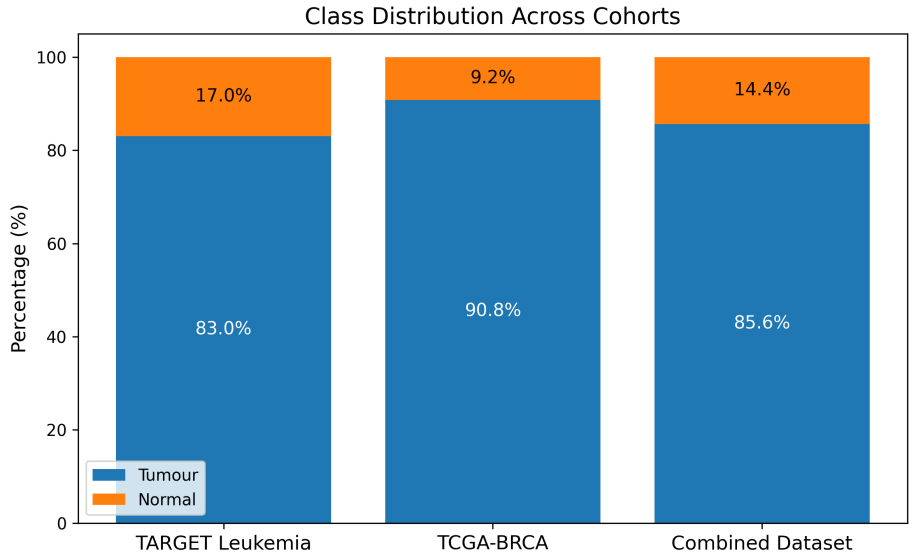


Figure 3.6: Class distribution across cohorts. Percentages of tumour and normal samples are shown for TARGET Leukemia, TCGA-BRCA, and the combined dataset. Total samples per Leukaemia data: 2374, Breast Cancer 1157 and for Combined 3531

3.2 PANDA Input Data and Prior Construction

GRN reconstruction using PANDA requires three primary inputs:

1. Gene expression matrix $X \in \mathbb{R}^{n_{genes} \times n_{samples}}$
2. Motif prior matrix $M \in \mathbb{R}^{n_{TF} \times n_{genes}}$
3. TF-TF PPI matrix $P \in \mathbb{R}^{n_{TF} \times n_{TF}}$

Gene identifiers were harmonised across datasets to ensure consistent mapping between expression genes, motif genes, and PPI nodes. The effective gene universe used for GRN reconstruction corresponds to the intersection:

$$G = G_{\text{expression}} \cap G_{\text{motif}} \cap G_{\text{PPI}}$$

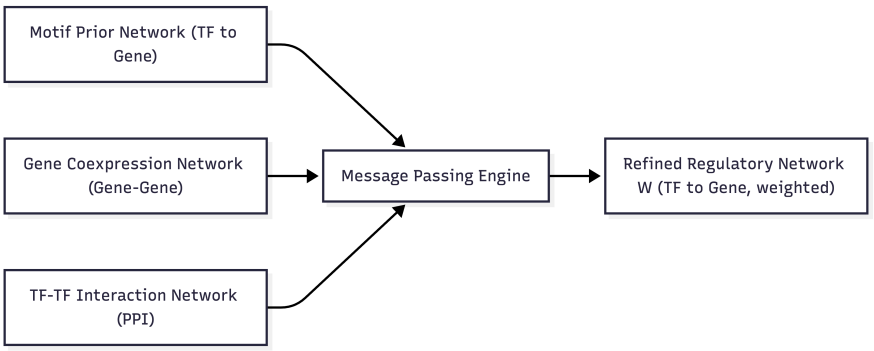


Figure 3.7: Schematic of PANDA inputs: motif prior (sequence-based or knowledge-based), TF-TF PPI network, and gene expression matrix jointly define the reconstructed GRN.

In this work, two alternative strategies were employed to construct the TF-gene prior, resulting in distinct GRN search spaces. These will be discussed in detail in section 3.4.2

3.2.1 Relation to LIONESS

LIONESS extends PANDA by estimating sample-specific networks through a leave-one-out strategy [48]. Since LIONESS operates directly on the aggregate GRN output of PANDA, no additional external input files are required beyond those described above. Differences in prior construction therefore propagate to the resulting patient-specific networks.

3.3 RNAseq Data Processing and Modeling Framework

3.3.1 Overview of the Modeling Strategy

This chapter describes the complete data processing and modeling framework adopted in this thesis. The overall strategy is divided into two complementary pipelines:

- **Expression-Centric Modeling Pipeline**
- **Regulatory Network–Driven Modeling Pipeline**

The first pipeline operates directly on RNAseq gene expression matrices and serves as a statistical and ML baseline. The second pipeline incorporates

biologically structured GRNs reconstructed using LIONESS and integrates them within advanced GNN architectures.

Dataset Acquisition

RNAseq data were obtained from the Genomic Data Commons (GDC) portal, focusing on the TARGET project (Therapeutically Applicable Research to Generate Effective Treatments), which provides pediatric leukemia cohorts including (ALL) and (AML).

The dataset has been filtered as described in chapter 3.1.2, Fig. 3.4 and contains:

- 2374 unique cases
- RNAseq expression (FPKM)
- Tumor and normal samples
- Rich clinical annotations

Raw Data Structure

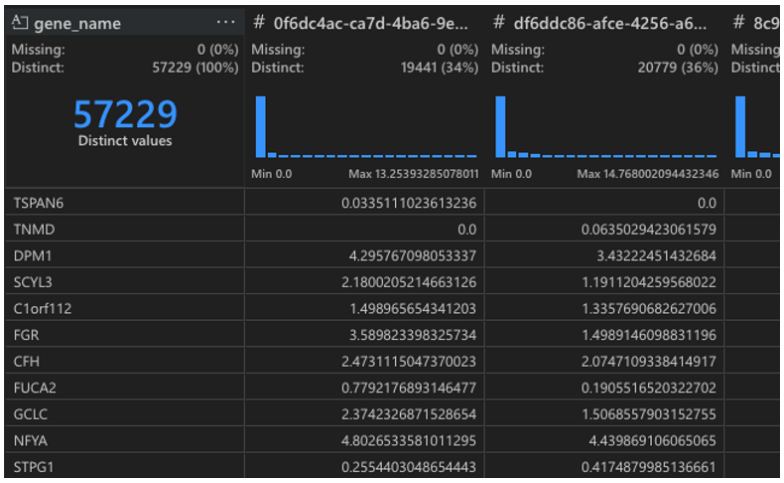
The RNAseq gene expression matrix can be formalized as:

$$\mathbf{X} \in \mathbb{R}^{G \times N} \tag{3.1}$$

where:

- $G \approx 60,000$ genes
- $N = 2374$ cases

In this work, expression values were handled in FPKM units (Fragments Per Kilobase per Million) and the pipeline was implemented to operate directly on the downloaded RNAseq data, supporting reproducibility across additional cohorts from the same portal [36]. A snippet of the matrix can be visualized in Fig. 3.8



gene_name	# 0f6dc4ac-ca7d-4ba6-9e...	# df6ddc86-afce-4256-a6...	# 8c9
Missing: 0 (0%) Distinct: 57229 (100%)	Missing: 0 (0%) Distinct: 19441 (34%)	Missing: 0 (0%) Distinct: 20779 (36%)	Missing Distinct
57229 Distinct values	Min 0.0 Max 13.25393285078011	Min 0.0 Max 14.768002094432346	Min 0.0
TSPAN6	0.0335111023613236	0.0	
TNMD	0.0	0.0635029423061579	
DPM1	4.295767098053337	3.43222451432684	
SCYL3	2.1800205214663126	1.1911204259568022	
C1orf112	1.498965654341203	1.3357690682627006	
FGR	3.589823398325734	1.4989146098831196	
CFH	2.4731115047370023	2.0747109338414917	
FUCA2	0.7792176893146477	0.1905516520322702	
GCLC	2.3742326871528654	1.5068557903152755	
NFYA	4.8026533581011295	4.439869106065065	
STPG1	0.2554403048654443	0.4174879985136661	

Figure 3.8: Snippet of the gene expression matrix. Raws correspond to genes and columns to unique cases.

3.3.2 Expression-Centric Modeling Pipeline

This section formalises the processing steps applied to RNAseq expression data before training expression-centric models. The workflow is aligned with the engineering pipeline summarised visible in the Fig. 3.9 and is designed to produce two downstream artefacts:

- a tabular feature matrix for classical ML/DL baselines (**Track A1**), and
- an expression-derived patient similarity graph for GNN modelling (**Track A2**).

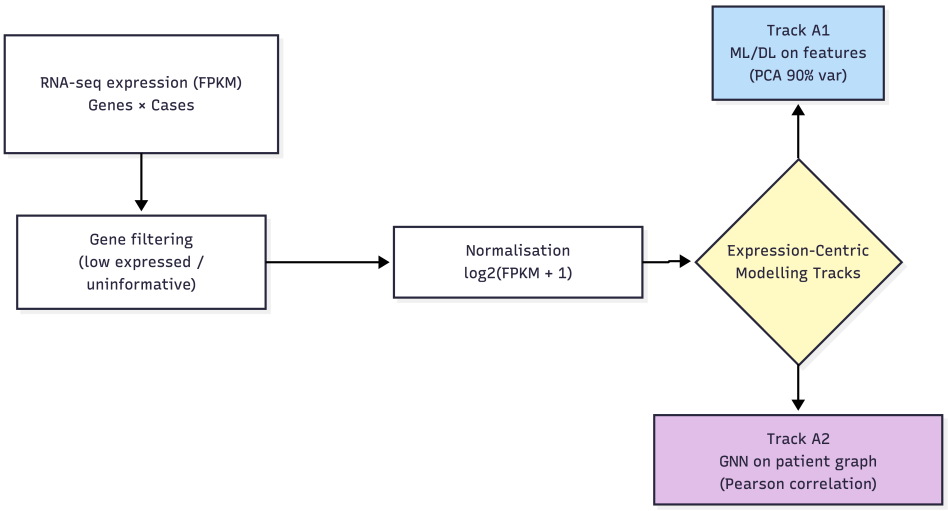


Figure 3.9: Overview of the Expression-Centric Modeling Pipeline.

Filtering and Quality Control (Expression-level)

An expression-level filtering step was applied to remove genes that are effectively uninformative across the cohort, a Seurat-like approach:

- low expressed: $FPKM > 0$ across all samples.
- uninformative: keep top-K high variance genes –> genes wich expressed the higher variance. Top-k values belong to [2000, 3000, 4000].

This reduces noise and computational overhead in downstream models. Seurat is primarily developed for single-cell RNAseq, its widely adopted analysis logic motivates the general principle of focusing on informative features (e.g., variable genes) and controlling low-signal dimensions before downstream learning [78, 85].

Normalisation and Transformation

Three preprocessing configurations were investigated in order to stabilise variance and reduce skewness in RNAseq expression values.

Raw expression values. In the baseline configuration, no transformation was applied and the input matrix corresponds directly to the measured expression values:

$$\mathbf{X}' = \mathbf{X} \tag{3.2}$$

Log-transformed expression values. To stabilise variance and reduce the impact of highly expressed genes, a $\log_2$ transformation was applied:

$$\mathbf{X}' = \log_2(\mathbf{FPKM} + 1) \tag{3.3}$$

Log-transformed expression with Z-score normalisation. In the third configuration, log-transformed values were further standardised using Z-score normalisation across samples. For each gene g, the expression values were normalised as:

$$Z_{g,i} = \frac{X'_{g,i} - \mu_g}{\sigma_g} \tag{3.4}$$

where μ_g and σ_g denote respectively the mean and standard deviation of gene g across all samples.

The final input matrix after this preprocessing step is therefore:

$$\mathbf{Z} = \frac{\log_2(\mathbf{FPKM} + 1) - \mu}{\sigma} \tag{3.5}$$

The *Log-transformed* transformation improves numerical conditioning and tends to benefit optimisation stability in ML/DL training. In the table 3.4 statistics of normalization and raw data are shown. Above Log normalization and Log + Z-score, the first is chosen because of the higher variance and instability of GNN models to negative expression.

Table 3.4: Comparison of normalization strategies applied to RNAseq expression data. The **$\mathbf{Log_2}$ transformation** (highlighted row) was selected as the normalization strategy adopted in this study.

Normalization Strategy	**Min**	**Max**	**Mean**	**Median**	**Std**
Raw Data	0	$4.6 * 10e6$	8.14	0.07	1085
$\mathbf{Log_2}$	**0**	**22.14**	**0.915**	**0.0975**	**1.45**
Log_2 + Z-Score	-11.7	62.98	1.87	-0.15	0.98

Track A1: Classical ML/DL on Expression Features

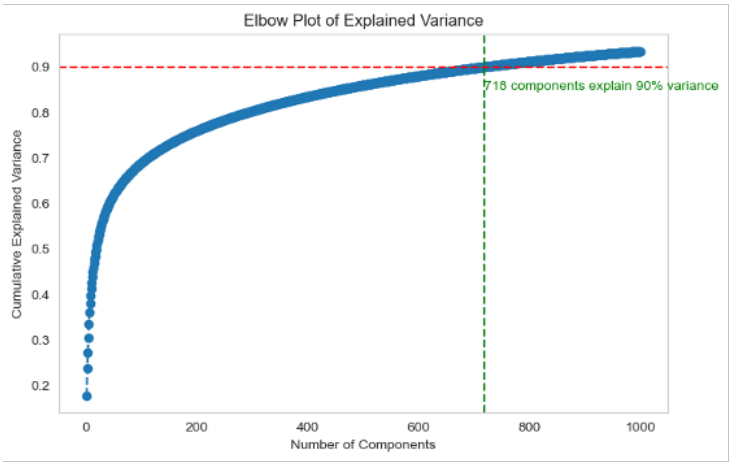
Track A1 models operate on sample-wise feature vectors derived from the normalized expression matrix $\mathbf{X}' \in \mathbb{R}^{G \times N}$, where G denotes the number of genes and N the number of samples.

Since classical ML models require observations arranged row-wise, the matrix was transposed to obtain a sample-by-gene representation:

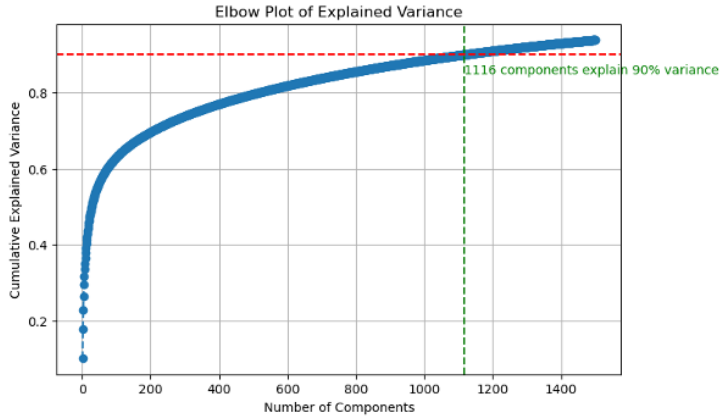
$$\mathbf{X}_{\mathrm{ML}} = (\mathbf{X}')^{\top} \in \mathbb{R}^{N \times G}. \tag{3.6}$$

In this representation, each row corresponds to a patient sample, and each column represents the expression level of a specific gene, thus treating gene expression values as features for supervised learning.

Dimensionality reduction was performed using PCA to retain a fixed fraction of variance (90% in the pipeline slides), producing a compressed representation with a reduced number of components/features. It's interesting to see that for raw data PCA components (718) are lower than for the normalized data (1116), mainly beacuse of the lower variance after normalization. Fig. 3.10b



(a) PCA on raw RNAseq data (n_components = 719)



(b) PCA on $\log_2$-normalized data (n_components = 1116)

Figure 3.10: Principal Component Analysis comparison between raw RNAseq data (left) and $\log_2$-normalized data (right). The normalized data exhibit improved variance stabilization and feature distribution prior to downstream modeling.

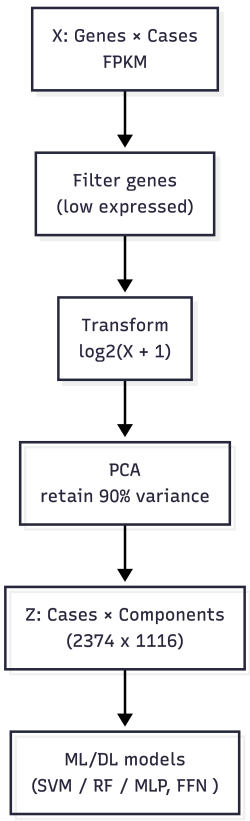


Figure 3.11: Track A1: preprocessing and feature construction for classical ML/DL models.

Track A2: GNN on Expression-Derived Patient Similarity Graph

Track A2 constructs a patient–patient similarity graph from the filtered expression matrix. Each node corresponds to a patient sample, and edges encode transcriptomic similarity computed from the (cases × genes) matrix after filtering and transformation. Following the workflow described in Fig. 3.12, the adjacency matrix was computed using Pearson cross-correlation:

$$A_{ij} = \mathrm{corr}(\mathbf{x}'_i, \mathbf{x}'_j), \tag{3.7}$$

where $\mathbf{x}'_i$ is the transformed expression profile of sample i. The resulting graph representation supports GNN-based learning in a purely expression-driven setting.

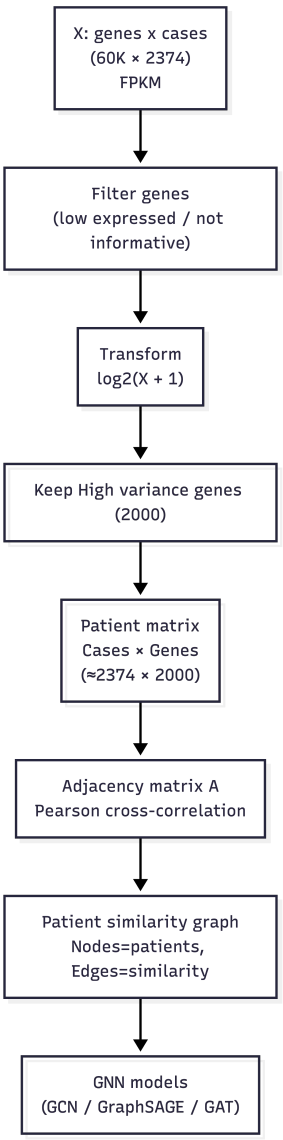


Figure 3.12: Track A2: patient similarity graph construction for expression-derived GNN modelling.

Selection of highly variable genes. To reduce dimensionality and focus the analysis on the most informative transcriptomic signals, highly variable genes (HVGs) were selected using a Seurat-inspired variance stabilization approach [14]. This method identifies genes whose variability across samples is higher than expected given their mean expression level.

Let $X \in \mathbb{R}^{G \times N}$ denote the gene expression matrix, where G represents genes and N samples. For each gene g, the mean expression μ_g and variance

σ_g^2 across samples are computed. A dispersion score is then defined as

$$d_g = \frac{\sigma_g^2}{\mu_g + \epsilon}$$

where ϵ is a small constant added for numerical stability.

Because gene variance strongly depends on expression magnitude, genes are first grouped into bins according to their mean expression using quantile-based binning. Within each bin, dispersion values are standardized by computing a normalized dispersion score:

$$\tilde{d}_g = \frac{d_g - \mathbb{E}[d_g]}{\mathrm{Std}(d_g)}$$

This normalization corrects for the mean-variance dependency and allows genes with unusually high variability to be identified regardless of their expression level. Finally, the top K genes with the highest normalized dispersion are selected as highly variable genes.

In this work, the top 2000 highly variable genes were retained and used as the feature space for downstream modeling.

Gene-gene similarity and adjacency matrix construction. To identify relationships between genes and construct graph structures suitable for GNNs, pairwise similarity between genes was computed using correlation-based measures. This step does not modify the dimensionality of the gene expression matrix but provides information about co-expression patterns between genes. In particular, the resulting similarity matrix is used to derive the adjacency structure required by GNNs, typically represented in PyTorch Geometric format as an edge index matrix of size $2 \times E$, where E denotes the number of graph edges.

Gene-gene relationships were primarily estimated using the Pearson correlation coefficient, which measures the linear dependency between two gene expression profiles across samples. Pearson correlation is widely used in transcriptomic analysis and forms the basis of gene co-expression network methods such as Weighted Gene Co-expression Network Analysis (WGCNA) [50]. Formally, for two genes g_i and g_j with expression vectors across samples, the Pearson correlation coefficient is defined as

$$r_{ij} = \frac{\sum_{k=1}^{N}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{N}(x_{ik} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{N}(x_{jk} - \bar{x}_j)^2}}$$

where x_{ik} represents the expression of gene i in sample k, and $\bar{x}_i$ denotes the mean expression of gene i.

In addition to Pearson correlation, two complementary similarity metrics were explored: cosine similarity and inverse Euclidean distance. Cosine similarity measures the angular similarity between expression vectors, while Euclidean distance captures absolute differences in expression magnitude. These measures provide alternative ways of quantifying transcriptomic similarity.

To compute these relationships, a custom function was implemented to generate similarity matrices and convert them into adjacency matrices. The function computes three pairwise similarity matrices (Pearson correlation, cosine similarity, and Euclidean-based similarity) from the gene expression matrix. Each similarity matrix is then thresholded to retain only sufficiently strong relationships, producing a sparse adjacency matrix that defines the graph connectivity.

The resulting adjacency matrices represent candidate gene-gene interaction structures that can be used to construct graph inputs for downstream GNN models.

Rationale of the Expression-Centric Approach

RNA sequencing (RNAseq) provides a high-resolution snapshot of the transcriptome and captures the functional state of the cell, reflecting the downstream outcome of multiple upstream regulatory mechanisms including transcriptional control, epigenetic regulation, and signaling pathways [10].

Several studies have demonstrated that gene expression profiles alone can achieve high classification performance in leukemia subtype prediction, particularly in distinguishing acute lymphoblastic leukemia (ALL) from acute myeloid leukemia (AML) using ML models applied to transcriptomic data [19, 31, 79].

However, conventional expression-based approaches typically treat genes as independent features and therefore do not explicitly model the regulatory interactions between transcription factors and their target genes. As a consequence, these methods cannot capture the complex regulatory network rewiring that characterizes oncogenic transformation and inter-patient heterogeneity in cancer [33, 48].

These limitations motivate the introduction of the *Regulatory Network Driven Modeling Pipeline*, described in the next section, which explicitly reconstructs GNNs to incorporate transcriptional control relationships into the learning framework.

3.4 GRN-based Dataset Construction and Graph Representation

This section describes the pipeline used to transform RNAseq gene expression data into LIONESS patient-specific GNNs (GRNs) and the subsequent strategies used to construct graph datasets for GNN (GNN) models.

The process is divided into two main parts:

1. Construction of patient-specific regulatory networks from gene expression data using the PANDA and LIONESS frameworks.
2. Graph generation strategies that convert the resulting GRNs into graph datasets suitable for GNN-based learning.

3.4.1 Pipeline from Gene Expression to LIONESS Networks

The GRN-based modelling approach starts from RNAseq gene expression profiles and reconstructs regulatory networks using the PANDA framework followed by LIONESS for patient-specific network estimation.

The overall pipeline is illustrated in Figure 3.13.

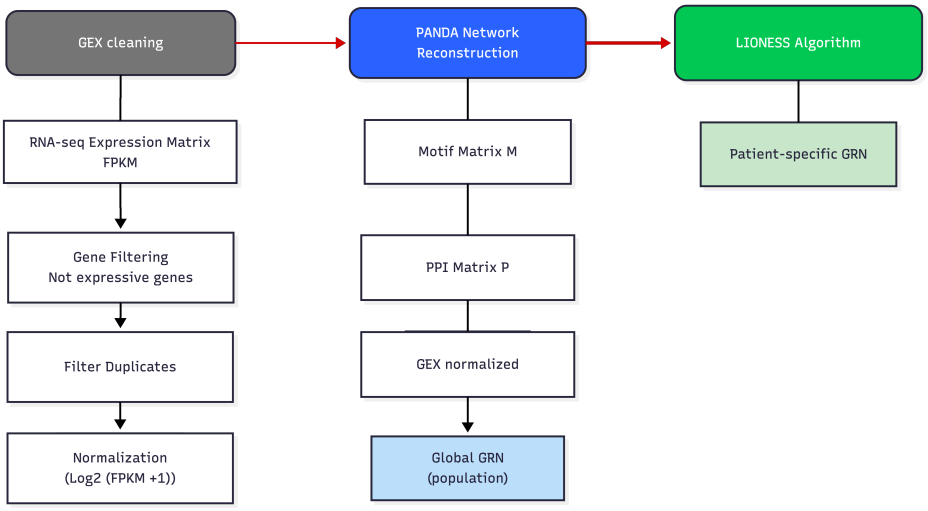


Figure 3.13: Pipeline from RNAseq gene expression to patient-specific LIONESS networks.

The process can be summarized as:

1. Gene expression preprocessing and filtering.

2. Construction of regulatory priors (motif and PPI matrices).

3. Reconstruction of a global GRN using the PANDA algorithm.

4. Generation of patient-specific GRNs using the LIONESS framework.

The PANDA and LIONESS algorithms are explained in chapter 2. PANDA integrates multiple biological sources to estimate regulatory interactions between transcription factors (TFs) and genes [33]. Subsequently, LIONESS extracts sample-specific networks from the global model [48].

3.4.2 Sequence-like and Knowledge-like Network Construction

Two different strategies were adopted to construct the regulatory priors used by PANDA:

- **Sequence-like networks**
- **Knowledge-like networks**

These two approaches differ in the level of prior biological knowledge provided to the PANDA algorithm.

Sequence-like GRN Construction

In the sequence-like strategy, regulatory interactions are inferred primarily from genomic sequence information. The PANDA algorithm therefore relies more strongly on its message-passing mechanism to infer regulatory relationships.

The process for building the motif prior and the ppi matrix is described in Fig, 3.14. The process can be resumed in (i) extracting the gene transcription start sites (TSS), (ii) identification of TF binding sites using TFLink annotations [1] and finally (iii) retrieval of protein-protein interactions downloaded from string.

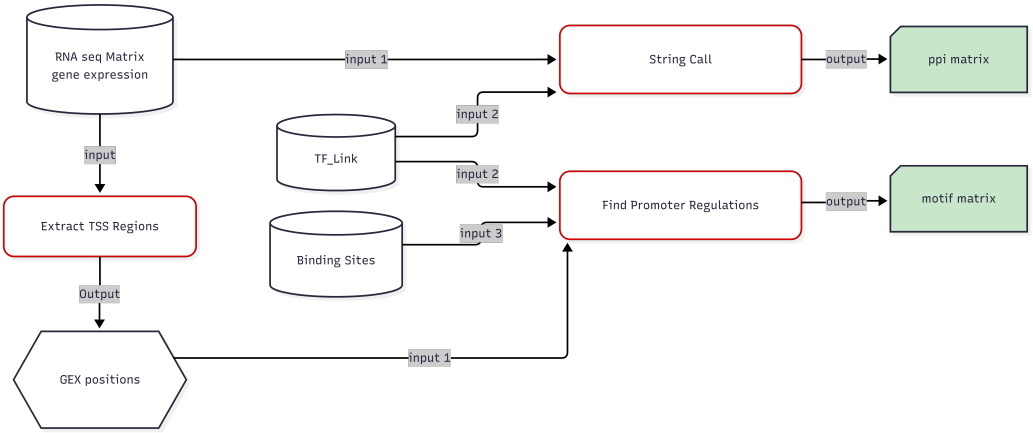


Figure 3.14: Generation of motif & ppi matrix for PANDA GRN creation - sequence strategy.

Knowledge-like GRN Construction

The knowledge-like strategy relies on curated regulatory interaction databases to define stronger biological priors for the GRN reconstruction process.

The main difference with respect to the sequence-like strategy lies in the binding-site interaction file employed [56], which contains experimentally supported TF-gene regulatory interactions and therefore provides a strong prior structure for the network reconstruction.

A second important difference concerns the procedure used to generate the motif and PPI matrices. In the sequence-like approach, gene coordinates are first used to identify promoter regions, which are subsequently scanned for transcription factor binding sites to construct the motif matrix. In contrast, the knowledge-like strategy directly filters curated TF-gene interaction data from RNAseq gex to construct the motif matrix without requiring promoter inference.

The overall process used to generate the PANDA input matrices for the knowledge-like configuration is illustrated in Figure 3.15.

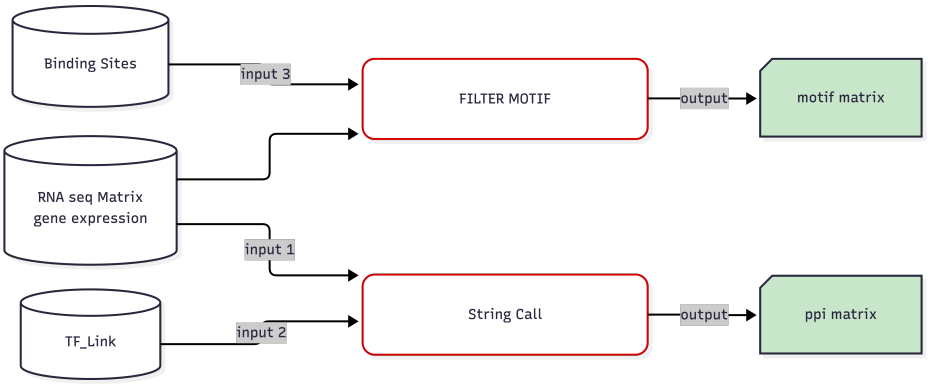


Figure 3.15: Generation of motif and PPI matrices used as input for PANDA GRN reconstruction in the knowledge-like strategy.

Under this configuration, PANDA starts from a stronger prior regulatory network and therefore focuses on refining existing regulatory relationships rather than inferring new interactions from sequence-based evidence.

Comparison Sequence and Knowledge strategies

Table 3.5: Comparison of regulatory prior statistics between the sequence-like and knowledge-like strategies used for PANDA GRN reconstruction. Leukemia dataset

Metric	**Sequence-like**	**Knowledge-like**
Motif matrix statistics		
Unique TFs	50	738
Unique genes	2,511	4,059
Possible TF-gene combinations	125,550	2,995,542
Actual TF-gene combinations (rows)	2,995	13,246
Genes that are also TFs	10	0
PPI matrix statistics		
Unique proteins in column 1	446	1,194
Unique proteins in column 2	451	5,554
Possible protein combinations	201,146	6,631,476
Actual protein combinations (rows)	1,208	24,058
Proteins appearing in both columns	251	958
Proteins not shared between columns	395	4,832

Table 3.5 reports summary statistics for the networks generated using both strategies. As expected, the knowledge-like configuration produces a larger

number of regulatory interactions as a result of the direct filtering process rather than using promoters previously generated from gene coordinates.

Together, these two complementary strategies allow the evaluation of two distinct GRN inference paradigms: a data-driven regulatory discovery approach (sequence-like) and a knowledge-driven network refinement approach (knowledge-like).

3.4.3 PANDA Network Reconstruction

For each prior configuration (sequence-like and knowledge-like), PANDA reconstructs regulatory networks using three different datasets:

- Tumor samples
- Normal samples
- Combined dataset (tumor + normal)

The PANDA output is a regulatory network:

$$W \in \mathbb{R}^{n_{TF} \times n_{genes}}$$

where each element represents the regulatory strength between a transcription factor and a gene. Here down the dataset analysed for Leukemia

Table 3.6: Graph dataset configurations

Dataset	Normalization	PANDA type
Dataset A	None	sequence-like
Dataset B	$log_2(FPKM + 1)$	sequence-like
Dataset C	$log_2(FPKM + 1)$	knowledge-like
Dataset D	$log_2(FPKM + 1) + Zscore$	knowledge-like

3.4.4 LIONESS Sample-Specific Networks

The LIONESS algorithm estimates patient-specific regulatory networks from the combined PANDA model.

Given a dataset with N samples, the LIONESS formulation computes the network of sample i as:

$$W^{(i)} = NW^{(all)} - (N-1)W^{(-i)}$$

where:

- $W^{(all)}$ is the network reconstructed using all samples

- $W^{(-i)}$ is the network reconstructed excluding sample i

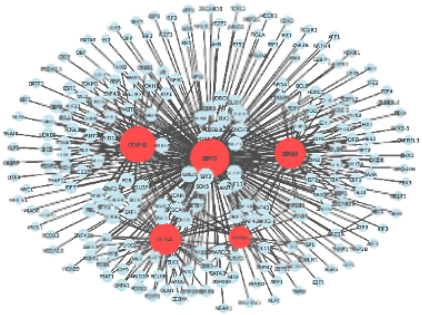
This procedure generates one GRN per patient.

Because the number of TF-gene interactions remains constant across networks, each LIONESS network contains approximately 6×10^6 regulatory edges.

Such networks are therefore extremely dense and require careful preprocessing before being used as input graphs for ML models.

tf	gene	motif	force
ZNF263	*IGLV1-47*	1.0	56.487034
NFAT5	*TNF*	1.0	56.387882
ZEB2	*CDH1*	1.0	56.256348
VDR	*BGLAP*	1.0	55.041286
FOXM1	*PDGFA*	1.0	54.688675
FOXA1	*PDCD6IPP1*	1.0	52.894047
NFE2L2	*ABCC1*	1.0	52.667629
MEF2A	*SLC2A4*	1.0	52.054054
MZF1	*MYB*	1.0	51.981480
POU2F1	*IL3*	1.0	51.599865

(a) Example snippet of a LIONESS-derived regulatory network.



(b) Graph density representation of the LIONESS network. 5 hubs and neighborhood.

Figure 3.16: Illustration of the structure and density characteristics of LIONESS-derived GNNs.

3.4.5 Graph Construction Strategies from LIONESS Networks

Once the patient-specific networks are generated, several strategies can be used to convert them into graphs suitable for GNN models.

Each patient network is transformed into a graph:

$$G = (V, E)$$

where:

- nodes V correspond to genes or transcription factors
- edges E correspond to regulatory interactions inferred by LIONESS (*force* value).

Different graph representations were explored depending on three main components: node features, edge features, and edge filtering strategies. These elements determine how the regulatory information extracted from the LIONESS networks is translated into graph structures suitable for GNN models.

- **Node features.** Several alternatives were considered for representing node attributes. The simplest representation consists of binary indicators (e.g., TF/gene flags). Alternatively, node features can incorporate biological information such as gene expression values. Another option used is to use the node features, derived from the regulatory interaction strengths estimated by LIONESS. For each gene node, two features are computed: the average outgoing regulatory strength (mean *force* toward its target genes) and the average incoming regulatory strength (mean *force* received from regulating transcription factors). These features summarize the regulatory influence of each node within the network.

- **Edge features.** Edge attributes correspond to the regulatory interaction strength provided by the PANDA/LIONESS framework. Specifically, each edge is associated with the *force* value representing the inferred regulatory relationship between a transcription factor and a target gene.

- **Edge filtering strategies.** Given the extremely dense nature of LIONESS-derived networks, filtering strategies are required to reduce graph density and remove weak interactions. In this work, edges with negative regulatory strength are removed by applying the constraint $force > 0$. This filtering step preserves biologically meaningful interactions while simplifying the resulting graph topology.

3.4.6 Final Unifying Section: Summary of Data Transformations

The modelling strategies explored in this work follow three complementary perspectives for representing transcriptomic information: expression-centric models, expression-derived graphs, and regulatory network representations.

These approaches progressively increase the level of biological structure incorporated into the learning process. While expression-centric models treat genes as independent features, graph-based approaches introduce relational structure between genes, and regulatory-network models explicitly encode transcription factor-gene regulatory interactions.

Table 3.7 summarizes the main data representations adopted throughout this study and the corresponding modelling paradigms.

Table 3.7: Summary of data representations and modelling paradigms explored in this work.

Phase	**Representation**	**Dimensionality**	**Model Type**
Expression-Centric	Gene expression matrix	$n_{\text{genes}} \times n_{\text{samples}}$	RF / SVM / MLP / FFN
Expression-Graph	Patient-patient similarity graph	$n_{\text{samples}} \times n_{\text{genes}}$	GNNs
Regulatory-Network	LIONESS single-patient GRN	$n_{\text{TF}} \times n_{\text{genes}}$	Advanced GNNs

These three modelling perspectives reflect an increasing degree of biological prior knowledge embedded in the learning process. In particular, regulatory-network models incorporate transcriptional regulation explicitly, allowing the investigation of disease-associated regulatory programs rather than relying solely on gene expression signals.

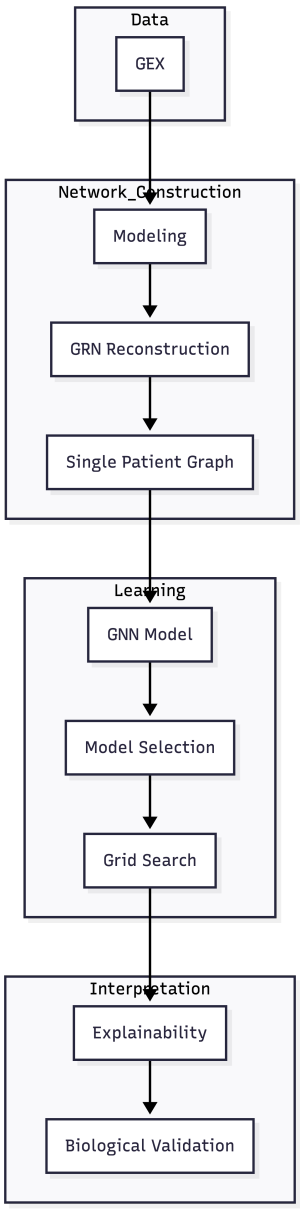


Figure 3.17: Overall pipeline of the proposed framework. RNAseq gene expression data are processed to reconstruct GNNs (GRNs), from which patient-specific graphs are generated. These graphs are used to train GNN models, followed by model selection and hyperparameter optimization. The final models are analysed using explainability methods and validated through biological interpretation.

The workflow summarized in Figure 3.17 integrates transcriptomic data

processing, regulatory network reconstruction, graph representation learning, and explainability analysis into a unified computational framework for leukemia subtype classification and cross-disease regulatory investigation.

Chapter 4

Methods and Modeling Framework

4.1 Overview of the Modeling Framework

This chapter describes the computational framework used to model transcriptomic and regulatory network data for leukaemia classification and cross-cancer disease association experiments. While Chapter 3 detailed the datasets and data processing steps that yield model-ready inputs, the present chapter focuses on the end-to-end modeling pipelines, including: (i) data representations adopted for learning, (ii) classical ML baselines, (iii) GNN architectures, (iv) training and validation strategies, (v) explainability methods, and (vi) experiment management and computational environment.

4.1.1 Modeling Pipelines and Data Representations

In this work, the same RNAseq data are leveraged through three complementary representations, leading to distinct modeling pipelines:

1. **Pipeline A - Gene expression vectors (baseline ML):** each sample is represented as a high-dimensional gene expression vector (after the processing described in Chapter 3). Classical ML models are trained on these vectors to provide a strong baseline.

2. **Pipeline B - Patient–patient similarity graphs (graph-based sample modeling):** samples are embedded into a graph where nodes correspond to patients and edges encode transcriptomic similarity (pearson correlation). GNNs are trained to classify patient nodes (or graph-level targets, depending on implementation).

3. **Pipeline C - Patient-specific regulatory networks (LIONESS GRN modeling):** each patient is associated with a personalised GNN inferred via PANDA and LIONESS. These patient-specific GRNs are encoded as graphs with genes as nodes and weighted regulatory edges, enabling GNNs to learn from structured regulatory patterns.

This explicit separation clarifies the experimental intent: to evaluate how different representations of transcriptomic information (vectors, patient graphs, regulatory graphs) affect predictive performance and interpretability within a unified experimental setting.

4.2 Transcriptomic Data Representations for Learning

This section formalises the three representations used for learning. Let $X \in \mathbb{R}^{G \times N}$ denote the processed gene expression matrix, where G is the number of genes and N the number of samples.

4.2.1 Representation A: Gene Expression Vectors

Each sample i is represented by a feature vector $\mathbf{x}_i \in \mathbb{R}^G$ (or in $\mathbb{R}^d$ if dimensionality reduction is applied, with $d \ll G$). This representation is the standard input for classical ML baselines and provides a reference point for performance comparisons.

Implementation Details: Full pipeline is shown in Fig.3.11. The following show the matrix dimension after each step:

- Initial dimensions: (genes x cases): (57235 x 2516)
- Drop sample duplicates: (genes x cases): (57235 x 2374)
- Filter not expressed genes: (genes x cases): (57229 x 2374)
- PCA retaining 90% of variance - binary problem: (cases x genes): (2374 x 1116)
- PCA retaining 90% of variance - multi-class problem: (cases x genes): (2374 x 858)

4.2.2 Representation B: Patient–Patient Similarity Graphs

A patient similarity graph is defined as $G^{(p)} = (V^{(p)}, E^{(p)})$, where $V^{(p)}$ is the set of patients and $E^{(p)}$ encodes similarity relationships. Node features correspond to transcriptomic vectors (original or reduced), while edges are constructed via a similarity rule (pearson cross-correlation).

Implementation Details: Full pipeline is shown in Fig. 2.4. The following show the matrix dimension after each step:

- Initial dimensions: (genes x cases): (57235 x 2516)
- Drop sample duplicates: (genes x cases): (57235 x 2374)
- Filter not expressed genes: (genes x cases): (57229 x 2374)
- Keep High variance genes: (cases x genes): (2374 x 2000).
- Pearson correlation: found 13669 graph connections.

4.2.3 Representation C: Patient-Specific GRNs from PANDA and LIONESS

For each patient i, a personalised regulatory network is inferred using PANDA as the population-level model, followed by LIONESS to estimate single-sample edge weights. The resulting patient-specific GRN is represented as a directed weighted graph $G^{(i)} = (V, E^{(i)})$, where V are genes (including TFs and targets) and $E^{(i)}$ are regulatory edges with weights ("force" scores). [33, 48]

The construction of these regulatory networks depends on the type of prior information used during the PANDA inference stage. Both, sequence and knowledge based strategies, are considered in this work.

Importantly, PANDA infers a global regulatory network whose topology defines the template structure for all subsequent LIONESS reconstructions. As a consequence, the dimensionality of the regulatory network remains constant across samples. In the case of the knowledge-based strategy, the inferred PANDA network comprises up to 28,000,089 transcription factor-gene interactions.

Each patient-specific regulatory network generated by LIONESS therefore represents a weighted instantiation of this shared network structure, where edge weights capture the contribution of the individual sample to the global regulatory model.

4.2.4 Comparison of transcriptomic data representations

The three representations introduced above capture different aspects of transcriptomic data and therefore lead to different modeling strategies. Table 4.1 summarizes their main characteristics.

Table 4.1: Comparison of transcriptomic data representations used in this work.

Representation	Nodes	Edges	Features	Model
Gene expression vectors	patients	none	normalized gene expression values	ML/DL
Patient similarity graph	patients	similarity edges	patient gene expression profiles	GNN
Patient-specific GRN	genes	TF-gene regulatory interactions	gene expression per gene node	GNN

These complementary representations enable a systematic investigation of how different structural assumptions about transcriptomic data affect model performance and interpretability. In particular, while vector-based approaches treat genes as independent predictors, graph-based models can exploit relational information between biological entities, either at the patient level or at the regulatory network level.

4.3 Cross-Cancer Cohort Construction and Sample Reduction

To evaluate the robustness of the proposed framework in a cross-cancer setting, the leukemia cohort was integrated with an independent breast cancer cohort. However, a direct combination of the two datasets was not appropriate because of their substantially different class distributions and cohort sizes. In the original leukemia dataset, malignant samples largely dominated the class distribution, with 2,136 diseased samples and 175 healthy samples. The breast cancer cohort contained 1,051 diseased samples and 106 healthy samples. This mismatch introduced two main issues: first, different within-dataset imbalance ratios across the two cancer types; second, an overrepresentation of leukemia samples, particularly in the malignant class, which

could bias the learning process toward leukemia-specific patterns rather than shared cross-cancer signals.

For this reason, the leukemia cohort was reduced through class-aware subsampling, using target class sizes derived from the breast cancer cohort. In the implemented setting, the reduced leukemia cohort retained 1051 malignant samples and 106 healthy samples, resulting in a more controlled and comparable class composition for the downstream joint analysis. This strategy was designed to mitigate class dominance, reduce the computational burden of cross-cancer training, and preserve meaningful biological structure in the selected subset.

Two alternative sampling strategies were investigated and shown in Fig. 4.2.

4.3.1 Centrality-based sampling.

The first strategy aims to select the most representative samples within each class. Samples are first embedded into a common low-dimensional space using global principal component analysis (PCA). Then, for each class independently, the centroid of the class distribution is computed, and the samples closest to that centroid are retained.

This approach favors prototypical observations and tends to reduce noise and extreme outliers. As a consequence, it yields a cleaner and more compact subset, although at the cost of a moderate reduction in within-class variance. The use of low-dimensional embeddings and neighborhood-based representations is consistent with common best practices for high-dimensional transcriptomic data analysis [58].

4.3.2 Diversity-based sampling.

The second strategy aims instead to preserve transcriptomic heterogeneity. As in the previous case, samples are first projected into a global PCA space. However, sample selection is then performed using a farthest-point strategy within each class: starting from an initial representative point, samples are iteratively selected so as to maximize their minimum distance from the already selected set. This procedure favors broad coverage of the class distribution and preserves biologically meaningful variability, making it suitable when the objective is to retain heterogeneous transcriptomic patterns rather than only central representatives. Diversity-preserving sketching and farthest-first strategies have been proposed as effective ways to summarize large biological datasets while maintaining coverage of the underlying structure [25, 40, 82].

Overall, the two reduction strategies reflect different assumptions about the desired structure of the reduced cohort. The centrality-based approach emphasizes representative and denoised samples, whereas the diversity-based approach emphasizes coverage and heterogeneity. Both were therefore retained for comparative evaluation in the cross-cancer experiments.

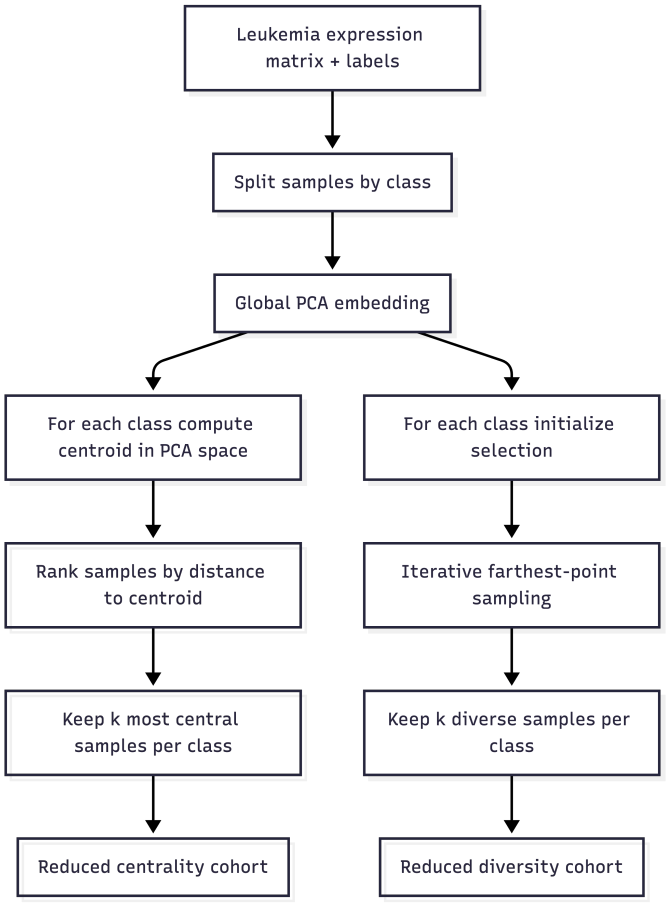


Figure 4.2: Conceptual illustration of the two cohort reduction strategies. Left branch: centrality-based sampling selects samples closest to the class centroid in PCA space, producing a representative and denoised subset. Right branch: diversity-based sampling uses farthest-point selection to maximize coverage of the transcriptomic space and preserve heterogeneity.

These two reduction strategies therefore allow evaluating the robustness of the proposed modeling framework with respect to different structural assumptions about the reduced cohort.

4.4 Classical ML Models

Classical ML models were used as baselines on the gene expression vector representation introduced in Section 4.2.1. Their role is twofold: first, to quantify the predictive signal already contained in transcriptomic vectors without explicit graph structure; second, to provide a reference for evaluating the added value of graph-based and regulatory-network-based models. Settings:

- The **feature matrix** used in this setting corresponds to the **PCA-transformed expression** representation, denoted by X_{PCA}.
- For all models, the dataset was **split** into **training and test** sets using a **stratified** hold-out strategy with an **80/20 proportion** and fixed random seed, thus preserving class proportions across the split.
- **Hyperparameter selection** was performed on the training set only using **randomized search with 3-fold cross-validation**.
- The **optimization** objective was the **macro-averaged F1 score**, chosen to reduce the impact of class imbalance by giving equal weight to all classes.

Formally, let $\mathcal{D}_{\text{train}}$ be the training set and $\mathcal{D}_{\text{test}}$ the test set. For each model m, the hyperparameter configuration θ_m^* was selected as

$$\theta_m^* = \arg\max_{\theta \in \Theta_m} \frac{1}{K} \sum_{k=1}^{K} \text{F1}_{\text{macro}}^{(k)}(\theta),$$

where Θ_m denotes the search space of model m, $K = 3$ is the number of cross-validation folds, and $\text{F1}_{\text{macro}}^{(k)}$ is the macro-F1 score obtained on fold k. The selected model was then retrained on the full training split and evaluated on the independent test split using balanced accuracy, macro-F1, classification report, confusion matrix, and training time.

Classical ML methods remain widely used in transcriptomic cancer classification and have also been applied to leukemia subtyping using gene expression profiles [79, 86].

4.4.1 Hyperparameter Optimization Strategy

Randomized hyperparameter search was adopted instead of exhaustive grid search in order to reduce computational cost while still exploring a sufficiently broad parameter space. For each model, 30 parameter configurations were

sampled. Table 4.2 and Fig. 4.3 summarize the search spaces used for the shallow baselines.

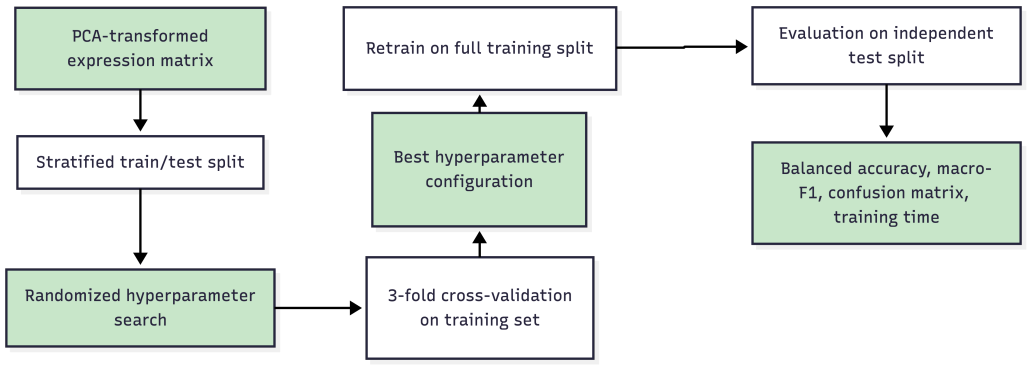


Figure 4.3: Scheme diagram of ML pipeline.

4.4.2 Summary of the Classical ML Baselines

Overall, the four baselines span complementary modeling assumptions on vectorized transcriptomic data: RF captures non-linear decision rules through tree ensembles; SVM provides strong margin-based classification in compact feature spaces; the optimized MLPClassifier introduces shallow neural modeling with hyperparameter search; and the custom feedforward neural network offers a manually controlled neural baseline. Together, these baselines establish a reference level of predictive performance before introducing graph-based relational inductive biases.

4.5 Graph Neural Network Architectures

GNNs are used in this work in two distinct contexts: (i) patient–patient graphs (Representation B), and (ii) patient-specific GRNs (Representation C). While the underlying GNN layers may be shared, the task formulation and graph semantics differ.

4.5.1 GNNs on Patient Similarity Graphs

In the second graph representation, Section 4.2.2, transcriptomic data are modeled as a patient-patient similarity network. In this graph, nodes correspond to patients and edges encode similarity relationships derived from gene expression profiles.

Table 4.2: Hyperparameter configurations explored for classical ML models.

Model	Hyperparameter	Values explored	Notes
RF	n_estimators	50, 100, 200	Number of trees
	max_depth	None, 10, 20	Maximum tree depth
	class_weight	balanced, balanced_subsample, None	Class imbalance handling
SVM	C	0.1, 1, 10, 100	Regularization strength
	kernel	linear, rbf, poly	Kernel function
	gamma	scale, auto	Kernel coefficient
MLPClassifier	hidden_layer_sizes	(64), (128), (128,64), (128,128)	Network architecture
	activation	tanh, relu	Activation function
	solver	sgd, adam	Optimizer
	alpha	$10^{-4}, 10^{-3}, 10^{-2}$	L_2 regularization
	learning_rate	constant, adaptive	Learning rate schedule

The graph is formally defined as:

$$G^{(p)} = (V^{(p)}, E^{(p)})$$

where $V^{(p)}$ represents the set of patients and $E^{(p)}$ the similarity edges connecting samples with similar transcriptomic profiles.

Input features. Each node is associated with a feature vector corresponding to the normalized gene expression representation described in Chapter 3.

Task formulation. The learning problem is formulated as a **node classification task**, where the objective is to predict the class label associated with each patient node. In this setting, the model does not operate directly

on gene interactions but rather learns patterns in the structure of the patient similarity network.

This formulation captures clusters of transcriptomically similar patients. As a consequence, explanations derived from this representation describe relationships between patients rather than the contribution of specific genes.

Although this approach may provide limited biological interpretability for single-disease classification, it becomes particularly relevant in the cross-cancer setting considered in this work. In that context, the similarity network integrates patients from different cancer types, allowing the model to identify transcriptomic neighborhoods shared across diseases. This perspective can highlight intermediate or atypical patients whose molecular profiles lie between multiple disease classes, potentially revealing shared oncogenic mechanisms.

Patient similarity networks have been widely used in biomedical ML to integrate heterogeneous patient data and improve disease classification through network-based learning [68, 103]. GNNs have recently been applied to such representations to model population-level relationships between patients [101, 104]. To ensure fair comparison between architectures, the same two-layer structure, activation functions, and dropout rate were used across all GNN variants (Fig. 4.4).

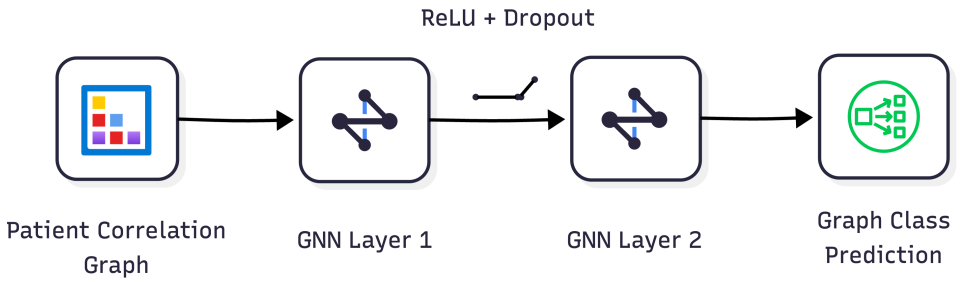


Figure 4.4: Generic architecture used for patient similarity graphs. Two graph GNN layers are applied with intermediate ReLU activation and dropout ($p = 0.5$). The GNN operator corresponds to GCN, GraphSAGE, or GAT depending on the experiment.

4.5.2 GNNs on Patient-Specific GRNs

The third representation considered in this work corresponds to patient-specific GNNs reconstructed through the PANDA and LIONESS framework described in section 3.4. In this setting, each patient is represented by an

individual regulatory graph inferred from the gene expression matrix and biological priors.

Formally, each sample i is associated with a graph

$$G^{(i)} = (V, E^{(i)})$$

where V denotes the set of genes and $E^{(i)}$ represents transcription factor-gene regulatory interactions inferred for patient i. Edge weights correspond to LIONESS regulatory strength scores, often referred to as *force* values.

Input features. Node features correspond to gene expression levels for the given patient. The resulting graph therefore integrates two complementary sources of information: the regulatory topology inferred from LIONESS and the patient-specific expression profile used as node attributes.

Task formulation. In this representation, the learning problem is formulated as a **graph classification task**. Each patient-specific GRN constitutes a separate graph labeled according to the clinical class of the patient (e.g., tumor vs normal or leukemia subtype). The GNN processes the graph structure and node features to predict the label associated with the entire network.

Unlike the patient similarity representation, this approach operates directly at the level of genes and regulatory interactions. As a result, explanations derived from the model can highlight important genes, regulatory edges, or subnetwork patterns associated with specific disease classes.

Edge directions inferred by LIONESS are treated as weighted interactions within the graph.

This representation forms the core modeling strategy of the present work. By combining patient-specific regulatory networks with graph neural architectures, the proposed framework aims to capture disease-specific regulatory programs while preserving biological interpretability.

Previous studies have shown that regulatory network representations can improve the interpretability of ML models in transcriptomic analysis by linking predictions to biologically meaningful interactions [11, 33, 48, 104].

Figure 4.5 illustrates the generic architecture adopted for GNNs applied to patient-specific GRN graphs. As in the patient-similarity setting, all models share the same two-layer GNN backbone with intermediate ReLU activation and fixed dropout probability $p = 0.5$. However, since the task is formulated as graph classification rather than node classification, an additional readout stage is required. After the second graph convolution block, node embeddings are passed through a second ReLU activation, aggregated by a

graph pooling operation, and finally mapped to the class space through a fully connected layer. Keeping the overall architecture fixed ensures a fair comparison between different GNN operators, while allowing each model to specialize through its own graph convolution mechanism.

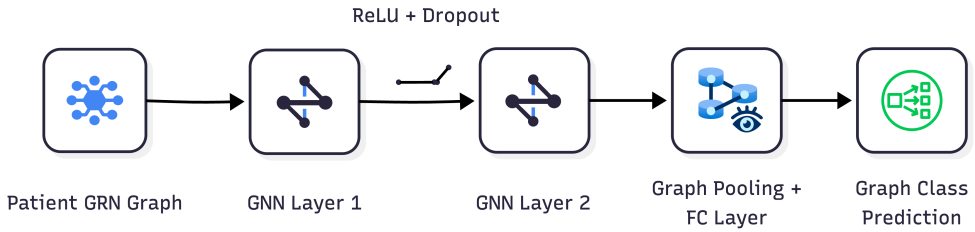


Figure 4.5: Generic GNN architecture used for patient-specific GRN graphs. The model consists of two graph convolution layers with ReLU activation and dropout ($p = 0.5$), followed by graph pooling and a fully connected layer for graph-level classification. The specific graph convolution operator corresponds to GCN, GraphSAGE, or GAT depending on the experiment.

4.5.3 Architectures Implemented

Graph Convolutional Network (GCN)

The Graph Convolutional Network (GCN) was implemented as one of the reference graph neural architectures for both patient similarity graphs (Representation B) and patient-specific GRN graphs (Representation C). To ensure a fair comparison across graph operators, the same overall two-layer backbone (Fig. 4.5) was adopted in both settings, while allowing selected hyperparameters to vary depending on the input representation.

The implementation details are visible in the project repository [29].

For Representation C, the graph encodes weighted regulatory interactions inferred from PANDA and LIONESS. Negative *force* values may cause undesirable behavior in the normalized propagation operator. For this reason, as discussed in Chapter 3, the GRN graphs were filtered so that only valid non-negative edge weights were retained before graph construction.

GraphSAGE

GraphSAGE was implemented as a second graph neural architecture to evaluate the robustness of the modeling pipeline across different neighborhood aggregation strategies. Unlike the spectral formulation of GCN, GraphSAGE

learns node embeddings by sampling and aggregating features from local neighborhoods through a parametric aggregation function [38].

In this work, GraphSAGE layers operate on the graph structure defined by `edge_index` and node feature matrices. In contrast to GCN, the implementation used does not directly incorporate edge weights or edge attributes within the convolution operator. Consequently, both patient similarity graphs (Representation B) and patient-specific GRN graphs (Representation C) are processed using the same connectivity structure without explicitly weighting edges during message passing.

Graph Attention Network (GAT)

The Graph Attention Network (GAT) was implemented as the third graph neural architecture considered in this work. Unlike GCN and GraphSAGE, GAT assigns learnable attention coefficients to neighboring nodes, allowing the model to weigh different neighborhood contributions adaptively during message passing [89].

In this work, the GAT architecture follows the same general two-layer backbone used for the other GNN models, with intermediate ReLU activation and dropout, in order to ensure a controlled comparison across graph operators.

The use of attention is particularly relevant in the GRN setting, where not all regulatory neighbors are expected to contribute equally to the node representation. In Representation C, this mechanism is further informed by regulatory edge attributes, allowing the model to incorporate not only graph topology but also the strength of inferred transcription factor–gene interactions. By contrast, in Representation B the attention mechanism operates exclusively on the patient similarity structure and the associated node features.

Other Implemented Architectures

In addition to the baseline GNN architectures described above, several additional models were implemented in order to explore alternative message passing mechanisms and assess their impact on graph classification performance. These models follow the same overall training pipeline and two-layer backbone described in Fig. 4.5. Only the graph convolution operator differs. A summary comparison table is shown in Table 4.3

GCN2. The GCN2 architecture extends the standard Graph Convolutional Network by introducing initial residual connections and identity mapping,

enabling deeper and more stable propagation of node features [17]. This design mitigates the over-smoothing problem typically observed in standard GCN layers.

Since the architecture differs from the standard GCN formulation, the forward propagation can be summarized as:

$$x \to x_0 \to \text{GCN2}(x, x_0) \to \text{GCN2}(x, x_0) \to \text{pool} \to \text{classifier}$$

where x_0 denotes the initial node representation used by all GCN2 layers to preserve the original feature information during propagation.

A key difference with respect to standard GCN is that each layer explicitly mixes the current node embedding with the initial feature representation, preserving informative signals from the original node features.

GATv2. The second version of the Graph Attention Network replaces the static attention mechanism used in `GATConv` with a more expressive dynamic attention formulation [13]. In contrast to the original GAT operator, the attention coefficients are computed after feature transformation, which increases the expressive power of the model and improves stability during training.

The architecture is the same of standard `GATConv`.

GIN. The Graph Isomorphism Network (GIN) uses multi-layer perceptrons to update node embeddings, providing a theoretically powerful message-passing scheme capable of distinguishing graph structures up to the Weisfeiler-Lehman test [92].

The implemented architecture includes:

- Two `GINConv` layers, each parameterized by a small MLP.
- ReLU activation and dropout between layers.
- Graph-level aggregation through `global_mean_pool`.
- Final linear layer for classification.

GINE. GINE extends the GIN architecture by incorporating edge attributes into the message passing process [42]. This capability is particularly relevant for GRN graphs, where regulatory interactions carry quantitative weights.

The implementation follows the same structure as GIN but replaces `GINConv` with `GINEConv`, enabling the use of edge features derived from LIONESS regulatory force scores.

GRNFormer. To further explore attention-based graph architectures, a transformer-style graph model was also implemented using `TransformerConv` from PyTorch Geometric. This architecture introduces attention-based message passing combined with edge-aware attention mechanisms.

The architecture consists of:

- Two stacked `TransformerConv` layers with multi-head attention - 4 heads.
- ReLU activation and dropout between layers (0.5).
- Graph-level readout through `global_mean_pool`.
- Final linear classifier.

Despite their architectural differences, all models share the same high-level pipeline consisting of two graph convolution layers, non-linear activation, dropout regularization, and graph-level pooling followed by a linear classification head. This design ensures a controlled comparison between alternative graph convolution operators.

Table 4.3: Comparison of additional GNN architectures implemented in this study.

Model	**Edge features**	**Attention**	**Residual / initial features**	**Core idea**
GCN2	optional edge weights	no	yes	deep GCN with initial residual connections
GATv2	optional edge features	yes	no	dynamic attention mechanism over neighbors
GIN	no	no	no	MLP-based aggregation with high expressive power
GINE	yes	no	no	GIN extended with edge attributes
GRNFormer	yes	yes	no	transformer-style attention message passing

Among the tested architectures, GCN2 is particularly suitable for regulatory networks because the initial residual connection preserves gene-level information while allowing deeper propagation of regulatory signals.

4.5.4 Readout and Classification Head

The final stage of the GNN pipeline depends on the prediction task associated with the graph representation.

Patient similarity graphs (Representation B). For patient similarity graphs, the task is formulated as node classification. Each node corresponds

to a patient, and the objective is to predict the class label of each node directly from its learned embedding. Consequently, the output of the second GNN layer is mapped directly to the class space without requiring a graph-level readout operation.

Patient-specific GRN graphs (Representation C). In contrast, the GRN-based representation corresponds to a graph classification problem, where each patient-specific regulatory network represents a single sample. After the second graph convolution layer, node embeddings must therefore be aggregated into a single graph-level representation.

This aggregation step is implemented through a global mean pooling operation provided by `torch_geometric.nn`:

$$\mathbf{h}_G = \frac{1}{|V|} \sum_{v \in V} \mathbf{h}_v$$

where $\mathbf{h}_v$ denotes the embedding of node v and $|V|$ is the number of nodes in the graph. The resulting graph-level embedding $\mathbf{h}_G$ is then passed to a fully connected linear layer that maps the pooled representation to the final prediction space.

This readout mechanism allows the model to summarize the global regulatory state of each patient-specific GNN while preserving the relational information captured during message passing.

4.6 Training Procedure and Hyperparameter Selection

This section describes the training procedures adopted for the GNN models. Since the training strategy depends on the underlying input representation, it is reported separately for patient similarity graphs (Representation B) and patient-specific GRN graphs (Representation C). The classical ML baselines operating on gene expression vectors (Representation A) were already described in Section 4.4, together with their randomized hyperparameter search spaces and cross-validation pipelines.

4.6.1 Training on Patient Similarity Graphs

For Representation B, all GNN models operate on a single patient–patient similarity graph, where nodes represent patients and node features correspond to the PCA-transformed transcriptomic profiles described in Chap-

ter 3. The learning task is node classification. Accordingly, training and evaluation are performed by splitting the node set into train and test folds while keeping the graph structure fixed.

In this setting, no additional hyperparameter grid search was performed. Instead, the training configuration was fixed across GCN, GraphSAGE, and GAT in order to ensure a controlled comparison between graph operators. This choice was motivated by the relatively limited scale of the patient graph and by the objective of isolating the effect of the message-passing mechanism rather than introducing extensive model-specific tuning. The adopted configuration is also consistent with common practice in shallow GNN node-classification benchmarks, where two-layer architectures, hidden dimensions around 64, Adam optimization, learning rate 0.01, weight decay 5×10^{-4}, and dropout around 0.5 are frequently used. In particular, the GAT literature commonly employs an 8-head first attention layer, while broader comparative benchmark studies report two-layer GNN backbones with hidden sizes in the 32–64 range and dropout values between 0.4 and 0.8 [38, 57, 80, 89].

Fixed training configuration. The parameters used for training the patient-similarity GNNs are summarized in Table 4.4.

Table 4.4: Fixed training configuration for GNNs on patient similarity graphs (Representation B).

Parameter	**Value**
Task	Node classification
Number of folds	5-fold stratified cross-validation
Epochs	100
Optimizer	Adam
Learning rate	0.01
Weight decay	5×10^{-4}
Hidden dimension	64
Dropout	0.5
Random seed	42
Loss function	Cross-entropy loss on training nodes
Evaluation metrics	Accuracy, balanced accuracy, macro precision, macro recall, macro F1

Training workflow. The training pipeline follows the same procedure for all GNN architectures and can be summarized as follows:

1. The full node-feature matrix X and the patient similarity graph con-

nectivity (`edge_index`) are loaded and converted to PyTorch tensors.

2. A stratified 5-fold split is generated over the patient labels using `StratifiedKFold`, ensuring that each fold preserves the class distribution.

3. For each fold, the model is initialized with the selected graph operator (GCN, GraphSAGE, or GAT) and trained for 100 epochs on the training nodes only.

4. At each epoch, the model computes node logits for the full graph, but the loss is evaluated only on the training indices of the current fold.

5. After training, the model is evaluated on the held-out test nodes of the same fold.

6. Performance is summarized through accuracy, balanced accuracy, macro precision, macro recall, and macro F1.

7. Fold-wise results are stored and finally averaged across all folds.

Since the graph is unique and shared across all patients, this training strategy corresponds to a full-graph transductive setting, where train and test nodes belong to the same graph but are masked differently across folds. Explainability artifacts are also saved at fold level in the implementation, but these outputs are discussed separately in Section 4.9.

Implementation note. The same training routine was used for all node-classification GNNs in Representation B. This design choice ensures that any observed performance differences can be attributed primarily to the graph convolution operator itself, rather than to changes in optimization or model-selection strategy.

Figure 4.6 illustrates the training procedure used for all GNN models operating on the patient similarity graph representation.

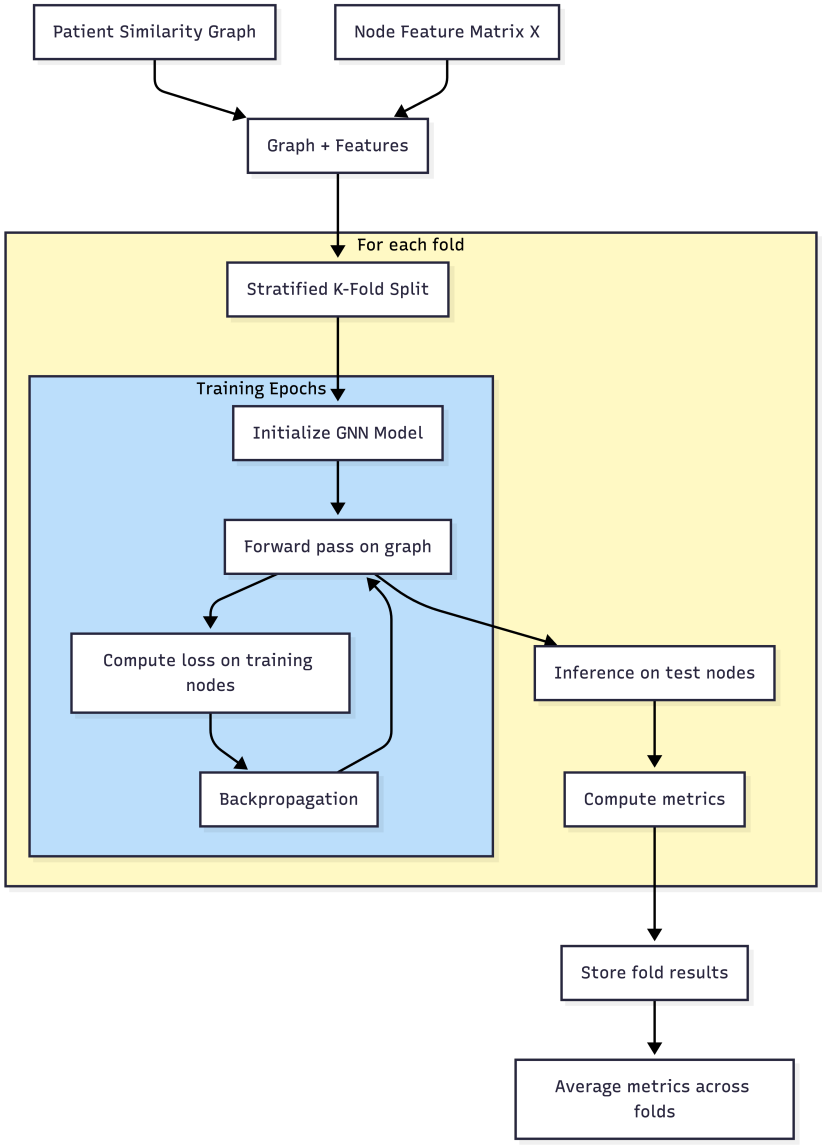


Figure 4.6: Training workflow for node classification on patient similarity graphs. A single graph is combined with node features and evaluated using stratified K-fold cross-validation. For each fold, the model is trained through multiple epochs and evaluated on held-out nodes.

4.6.2 Training on Patient-Specific GRN Graphs

For Representation C, the learning problem is formulated as graph classification. Each patient is associated with one patient-specific GRN reconstructed from the PANDA–LIONESS pipeline described in Chapter 3, and the objec-

tive is to predict the class label of the whole graph. Unlike Representation B, the training process therefore operates on a dataset of graphs rather than on a single shared graph with multiple labeled nodes.

Hyperparameter optimization for this setting was performed separately and is described in Section 4.7. The present subsection focuses only on the training procedure.

Graph construction variants. Before training, each patient-specific graph is instantiated according to one of several graph-construction modes, corresponding to alternative definitions of node features and edge weights. These settings define the graph input used in the downstream classification experiments and correspond to the different test configurations summarized in Table 4.5.

Table 4.5: Graph construction modes used for patient-specific GRN classification experiments.

Test	**Node features**	**Edges Transformation or filtering**
0	Dummy	force > 0
1	Dummy	force > 0 + z-score
2	[out-degree, in-degree]	force > 0
3	Gene expression ($\log_2$)	force > 0 + z-score
4	Gene expression ($\log_2$)	force > 0

In all cases, the graph topology is derived from the patient-specific LIONESS regulatory network. The differences between the test settings lie in the node representation and in the transformation applied to the edge weights. In particular, edge weights can be used either as positive raw force values or as normalized force z-scores. Likewise, node features can range from dummy vectors to simple structural descriptors or to patient-specific gene expression values. These variants were introduced to assess how predictive performance changes as the amount of biological information encoded in the graph increases.

Cross-validation strategy. Model evaluation is performed through stratified K-fold cross-validation with $K = 5$, using graph labels for stratification. At each outer fold, the graph dataset is first split into training and test partitions. The training partition is then further divided into training and validation subsets using a stratified hold-out split. This validation subset is used for model selection during training and for early stopping.

Training procedure. The training process is shared across all graph-classification GNN architectures and is summarized in Fig. 4.7

Training configuration. The implementation used in this work adopts the following general training setup for Representation C: 5-fold cross-validation, 20% validation split inside each training fold, graph-level mini-batching with `batch_size = 1`, gradient accumulation over 4 steps, Adam or AdamW optimization, and early stopping based on validation balanced accuracy. The exact values of the architecture-dependent hyperparameters (e.g., hidden dimension, dropout, learning rate, weight decay, loss type, and edge-weight mode) are reported separately in the hyperparameter selection subsection.

Scope note. The training routine also saves fold-level artifacts required for downstream explainability analyses. However, those steps are not discussed here and are deferred to Section 4.9, where the explainability workflow is described in detail.

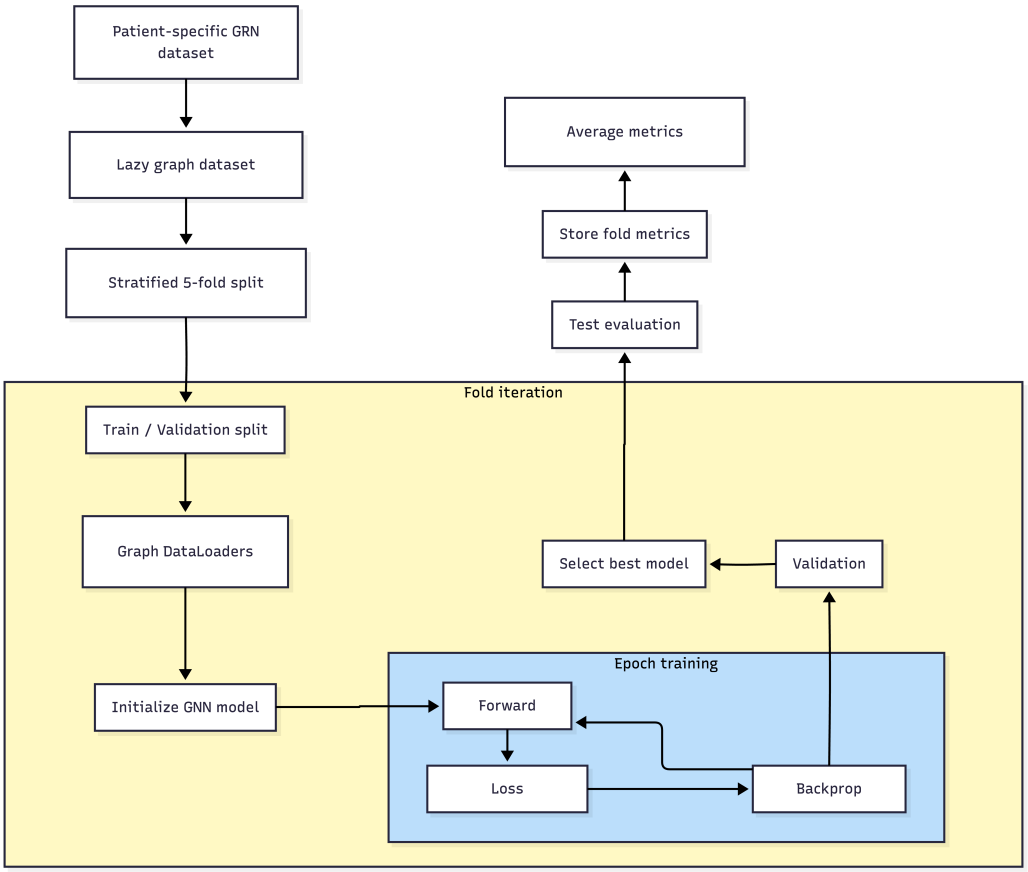


Figure 4.7: Training workflow for graph classification on patient-specific GRN graphs. Each patient graph is loaded on demand, split through stratified cross-validation, trained with graph-level supervision, validated for model selection, and finally evaluated on the held-out fold.

4.7 Hyperparameter Search

Hyperparameter optimisation was performed exclusively for the models operating on the patient-specific GRN representation (Representation C). Due to the high computational cost of training GNNs on large regulatory graphs, the search procedure was organised in a multi-stage strategy aimed at progressively identifying the most promising architectures and configurations.

In particular, GRN graphs reconstructed through the PANDA–LIONESS pipeline can contain tens of millions of potential regulatory interactions when derived from knowledge-based priors. Training a single model on such graphs may require several hours to up to 3-4 days per fold depending on the con-

figuration and hardware resources available. Consequently, an exhaustive hyperparameter exploration across all architectures was computationally infeasible.

To address this limitation, the search procedure was divided into three sequential phases:

1. **Model screening phase**: evaluation of the main baseline architectures with a moderate hyperparameter grid in order to identify the most promising model family.

2. **Architecture exploration phase**: evaluation of additional GNN architectures using the best-performing hyperparameter region identified in the previous phase.

3. **Final Three-Stage Hyperparameter Search** on the candidate model.

This hierarchical strategy allows efficient exploration of the model space while keeping the computational requirements manageable.

4.7.1 Phase I: Model screening phase

The first phase focuses on the baseline architectures introduced previously (GCN, GraphSAGE and GAT). A structured hyperparameter grid search was performed to explore the influence of key architectural and optimisation parameters. For computational efficiency, this stage uses a reduced cross-validation scheme with $n\_folds = 3$.

Training is performed using the Adam optimiser with cross-entropy loss. Early stopping is implemented through a patience counter to prevent unnecessary training when the model does not improve.

Table 4.6 summarises the hyperparameter search space explored in this stage.

Table 4.6: Hyperparameter search space for baseline GNN architectures (Phase I).

Hyperparameter	**Search space**
Hidden dimension	$\{16, 32, 64, 128\}$
Learning rate	$\{10^{-3}, 10^{-4}, 5 \times 10^{-4}\}$
Weight decay	$\{0, 5 \times 10^{-4}\}$
Epochs	$\{30, 50, 60\}$
Batch size	$\{1, 2, 4, 6, 8\}$
Number of folds	3
Optimizer	Adam
Loss function	CrossEntropy
Class weighting	{none, inverse frequency}
Early stopping patience	13 epochs
Seed	42
Number of workers	$\{1, 2, 4, 8\}$

The selected hyperparameters reflect both model expressiveness and practical constraints related to GPU memory and training throughput. In particular, batch size and number of workers mainly influence the computational performance rather than the model capacity.

The evaluation metric used to compare configurations is the **balanced accuracy**, which is particularly suitable for imbalanced biomedical datasets.

4.7.2 Phase II: Architecture exploration phase

After identifying the most promising hyperparameter region in Phase I, the search was extended to additional architectures including GCN2, GAT2, GIN, GINE and GRNFormer.

In this phase the hyperparameters are largely fixed to the best-performing values identified previously in order to isolate the contribution of the architecture itself. The goal of this step is therefore to compare different message-passing mechanisms rather than perform full hyperparameter optimisation.

The configurations adopted for this stage are summarised in Table 4.7.

Table 4.7: Training configuration used for extended architecture comparison (Phase II).

Parameter	**Value**
Epochs	30
Learning rate	3×10^{-4}
Weight decay	10^{-3}
Batch size	1
Gradient accumulation steps	1
Validation split	20%
Optimizer	AdamW
Loss function	CrossEntropy
Class weighting	balanced
Early stopping patience	8
Seed	42

The training procedure in this phase relies on a stratified train/validation split rather than cross-validation in order to significantly reduce computation time. Class imbalance is handled through weighted cross-entropy, where weights are computed according to the class frequency within the training subset.

This two-stage optimisation strategy allows efficient identification of the most suitable architecture for GRN-based classification while maintaining tractable computational requirements.

4.7.3 Phase III: Final Three-Stage Hyperparameter Search on the Candidate Model

After the architecture screening phases described above, a final hyperparameter search was performed on the selected model family. The objective of this stage was not to compare different graph operators, but to refine the training configuration of the most promising architecture under realistic computational constraints.

Because each training run on patient-specific GRN graphs is expensive, the final search was implemented as a three-stage successive filtering procedure. The same train/validation split was kept fixed throughout the search, while the training budget was progressively increased across stages. Candidate configurations were ranked according to validation balanced accuracy, and only the best-performing configurations were promoted to the next stage.

Search Configuration

The hyperparameter search is controlled through a dedicated configuration object, `SearchConfig`, which defines both the fixed training controls and the stage-wise search schedule. The main purpose of this design is to separate the general training constraints (e.g., batch size, accumulation steps, number of workers, validation split) from the parameters actually explored during the search.

The stage-wise search schedule is summarized in Table 4.8.

Table 4.8: Stage-wise search schedule for the final GCN2 hyperparameter optimisation.

Stage	Training budget	Candidates retained	Purpose
Stage 1	8 epochs	top 8	broad screening
Stage 2	20 epochs	top 3	refinement
Stage 3	40 epochs	top 1	final selection

Search Space

The final search space was defined through the `sample_trial` logic and the corresponding stage-wise grid generation. In the development script, a reduced subset of values was temporarily used to accelerate experimentation; however, the complete search space considered for the methodological description is the full set of candidate values originally defined in the code.

Table 4.9 summarizes the explored hyperparameters.

Table 4.9: Hyperparameter space explored in the final GCN2 search.

Hyperparameter	**Values explored**
Hidden dimension	$\{32, 64, 128\}$
Dropout	$\{0.2, 0.4, 0.6\}$
Learning rate	$\{2 \times 10^{-4}, 5 \times 10^{-4}\}$
Weight decay	$\{0, 5 \times 10^{-4}, 10^{-3}\}$
Optimizer	{`Adam`, `AdamW`}
Class weight mode	{`none`, `inverse`, `balanced`, `effective_num`}
Effective number parameter β	$\{0.99, 0.999\}$
Loss function	{`CrossEntropy`, `FocalLoss`}
Label smoothing	$\{0.0, 0.05\}$
Focal loss γ	$\{1.0, 2.0\}$
Gradient clipping	$\{0.0, 1.0, 2.0\}$
Edge-weight transformation	{`none`, `abs`, `clip0_5`, `signed_shift`}

In Stage 1, the full candidate grid is generated and then deterministically shuffled. To keep the first stage computationally tractable, at most 24 candidate configurations are evaluated. The resulting ranking is based on the best validation balanced accuracy reached during training. The top 8 configurations are then re-trained in Stage 2 with a longer budget, and the top 3 of these are finally evaluated in Stage 3 with the largest training budget.

Class Weighting Strategies

Because the class distribution of the training subset may be imbalanced, weighted losses are used during the search. Class weights are computed only on the training partition and can be selected through four modes:

- **none**: all classes receive equal weight;
- **inverse**: class weights are proportional to the inverse of the class frequency and normalized to unit mean;
- **balanced**: weights are computed as

$$w_c = \frac{N}{C\, n_c},$$

  where N is the total number of training samples, C the number of classes, and n_c the number of samples in class c;

- **effective_num**: weights are computed according to the effective number of samples formulation proposed by Cui et al. [23].

This design allows evaluating whether standard inverse-frequency reweighting or more refined imbalance-aware strategies improve robustness on graph-level classification.

Loss Functions

Two loss families are considered in the final search:

- **Weighted cross-entropy loss**, optionally combined with label smoothing;
- **Focal loss**, which down-weights easy examples and focuses the optimization on harder samples [55].

For weighted cross-entropy, the class weights described above are directly passed to the loss function. Label smoothing can optionally be enabled with smoothing coefficient 0.05.

For focal loss, the loss for a training sample is modulated by the factor $(1-p_t)^\gamma$, where p_t is the predicted probability of the true class and γ controls the focusing strength. In the present search, $\gamma \in \{1.0, 2.0\}$.

Edge-Weight Transformations

The candidate model can directly exploit scalar edge weights during message passing. For this reason, the final search also includes different transformations of the original GRN edge weights at training time:

- **none**: no explicit edge weights are passed to the model;
- **abs**: the absolute value of the edge weights is used;
- **clip0_5**: edge weights are clipped to the interval $[0, 5]$. The lower value (0) is related to filtering of poor TF-gene interactions, while the maximum value (5) is related to the distribution of the *force* value in all LIONESS graphs. $99th - percentile < 4$ for all normalized datasets studied ;
- **signed_shift**: all edge weights are shifted by subtracting the minimum value, producing a non-negative range.

These transformations were introduced to assess how the numerical treatment of regulatory edge weights affects optimization stability and predictive performance.

Three-Stage Search Procedure

The full search procedure can be summarized as follows:

1. The lazy GRN dataset is built once and graph labels are extracted.
2. A single stratified train/validation split is generated and kept fixed across all stages.
3. In **Stage 1**, up to 24 shuffled candidate configurations are evaluated for 8 epochs each.
4. Candidate configurations are ranked by best validation balanced accuracy, and the top 8 are retained.
5. In **Stage 2**, the top 8 configurations are re-evaluated for 20 epochs each.
6. The resulting configurations are again ranked, and the top 3 are retained.
7. In **Stage 3**, the top 3 configurations are re-evaluated for 40 epochs each.
8. The final best configuration is selected according to validation balanced accuracy and stored as the output of the search.

This design corresponds to a coarse-to-fine search: the first stage filters the search space broadly, the second stage refines the ranking among competitive candidates, and the third stage confirms the final best configuration using a larger training budget. Visual reference is represented in Fig. 4.8

The final best configuration follow then the training/validation procedure with 5 folds explained in section 4.6.2

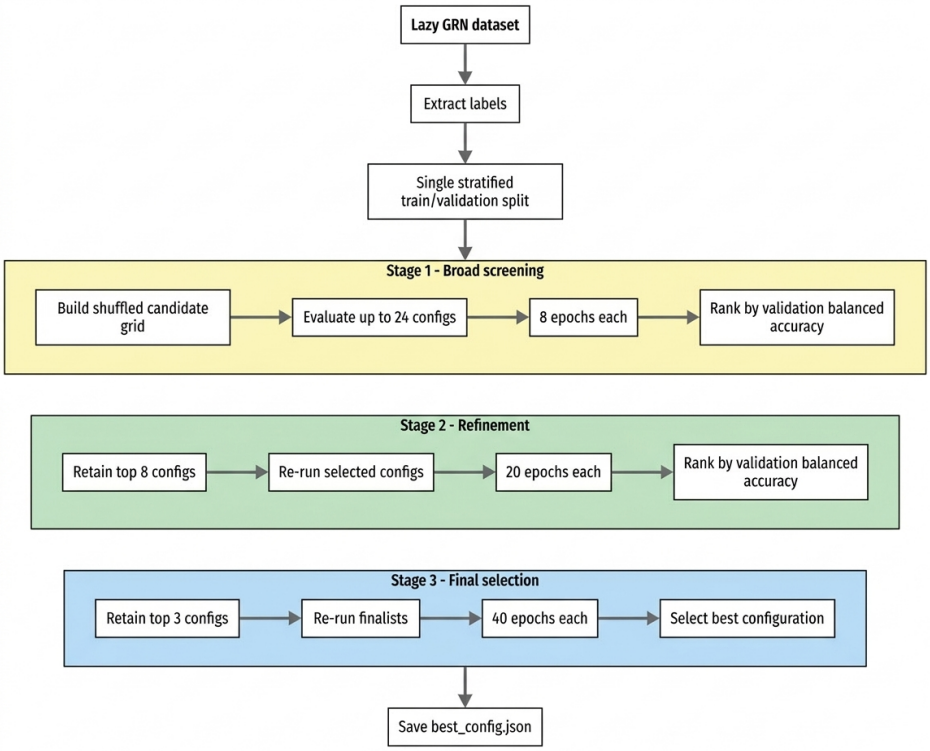


Figure 4.8: Three-stage hyperparameter search procedure used for the final GCN2 optimisation. Candidate configurations are progressively filtered according to validation balanced accuracy while the training budget increases across stages.

4.8 Dataset Splitting and Validation Strategy

This section describes the validation strategy adopted for GNNs operating on patient-specific GRN graphs reconstructed through the PANDA–LIONESS pipeline. Because training on large regulatory graphs is computationally expensive, model development was organised as a progressive multi-phase procedure. The validation protocol was therefore adapted to the objective of each phase: broad model screening, architecture exploration, final hyperparameter refinement, and faithful re-evaluation of the selected model.

Table 4.10 summarizes the four validation settings used throughout the study.

Table 4.10: Validation protocols used for GNNs on patient-specific GRN graphs.

Stage	**Objective**	**Split protocol**	**Selection criterion**
Phase I	baseline model screening	stratified 3-fold CV	mean balanced accuracy across folds
Phase II	architecture exploration	single stratified 80/20 split	best validation balanced accuracy
Phase III	exhaustive search on final candidate	single stratified 80/20 split	best validation balanced accuracy
Final evaluation	faithful assessment of best model	outer stratified 5-fold CV + inner 80/20 train/val split	average test performance across folds

4.8.1 Train/Validation/Test Splits

All experiments are performed at the **patient level**, where each graph corresponds to one patient-specific GRN and each label is associated with the entire graph. Splitting is therefore patient-independent by construction, which prevents information leakage across train and evaluation subsets.

Two main splitting strategies were used.

Single stratified split. In Phases II and III, a single stratified split is created once and reused for all candidate models/configurations within the phase. Specifically, the graph dataset is partitioned into:

- 80% training graphs,
- 20% validation graphs.

Nested train/validation split inside cross-validation. In the final faithful evaluation, the dataset is first split through outer stratified 5-fold cross-validation. Then, within each outer training fold, an additional stratified 80/20 split is applied to create:

- an inner training subset used for optimisation,

- an inner validation subset used for model selection and early stopping.

The held-out outer fold is used exclusively for final testing.
Table 4.11 summarizes these splitting schemes.

Table 4.11: Train/validation/test split strategies for GRN-based GNN experiments.

Stage	**Train**	**Validation/Test**	**Notes**
Phase I	fold-dependent	held-out fold	3-fold CV, no separate validation split
Phase II	80%	20% validation	fixed split reused across models
Phase III	80%	20% validation	fixed split reused across trials
Final evaluation	outer train fold	inner 20% validation + outer test fold	5-fold CV, faithful nested evaluation

4.8.2 Cross-Validation

Cross-validation is used in two distinct contexts:

1. Baseline model screening, the initial model screening stage uses **stratified 3-fold cross-validation**

2. Final faithful evaluation, after the best final configuration is selected, the model is re-evaluated through a **faithful outer 5-fold cross-validation**. In this setting, each outer training fold is further split into train/validation subsets, mimicking the optimisation regime used during hyperparameter search while preserving a clean held-out test fold for final assessment

4.8.3 Evaluation Metrics

The primary model-selection metric throughout the GRN-based GNN experiments is **balanced accuracy**. This choice is motivated by the class imbalance present in several experimental settings and by the need to evaluate performance symmetrically across classes.

Depending on the experimental phase, the following metrics are recorded:

- **accuracy**,
- **precision**,
- **recall**,
- **F1 score**,
- **balanced accuracy**.

Table 4.12 summarizes the role of the evaluation metrics in each stage.

Table 4.12: Role of evaluation metrics across validation stages.

Stage	**Metrics recorded**	**Main use**
Phase I	accuracy, precision, recall, F1, balanced accuracy	model screening across folds
Phase II	training loss, validation balanced accuracy, validation F1	architecture comparison
Phase III	training loss, validation balanced accuracy, validation F1	hyperparameter ranking
Final evaluation	accuracy, precision, recall, F1, balanced accuracy	final test assessment

Balanced accuracy is therefore used as the main ranking criterion at all critical decision points, whereas the other metrics are reported to provide a more complete description of model behaviour.

4.9 Explainability Methods

To complement predictive performance with biologically interpretable outputs, two explainability methods were applied to the trained GNN models on patient-specific GRN graphs: a *SHAP-like* attribution method based on sampled Shapley values, and *GNNExplainer*. The two methods were selected because they provide complementary views of the prediction mechanism. The first focuses primarily on **feature-level importance**, highlighting which

genes contribute most strongly to the model output, whereas the second focuses on **structure-aware explanations**, identifying relevant nodes and regulatory subgraphs. The selected explainers support both gene-level interpretation and network-level interpretation [102]

The explainability analysis is performed only on the GRN-based graph classification setting (Representation C), where each graph corresponds to one patient-specific regulatory network inferred through PANDA and LIONESS.

4.9.1 SHAP-like Attribution

In this work, SHAP-style explanations are implemented through `Shapley Value Sampling` [59] from the Captum library rather than through the standard SHAP package. For this reason, the method is referred to as *SHAP-like* rather than exact SHAP.

This choice is motivated by the nature of the input data. Standard SHAP implementations are primarily designed for flat tabular or tensor-based inputs, whereas the present study operates on graph-structured data with message-passing layers, where the model input depends jointly on node features, graph connectivity, and graph weights. In this context, the semantics of masking become less straightforward, since one may wish to mask features, nodes, or edges. By contrast, Captum provides a direct perturbation-based approximation of Shapley values within the native PyTorch framework, making it easier to define a custom forward function for graph models.

In the implemented pipeline, SHAP-like attribution is computed as follows:

- each patient-specific GRN is processed independently (`batch_size = 1`);
- the explainer operates on the node feature matrix x, while the graph structure is kept fixed;
- the baseline is defined as a zero tensor with the same shape as x;
- the target output is the logit associated with the true class of the graph;
- the resulting attribution tensor is reduced to a single score per node by taking the absolute value and summing over the feature dimension.

Since in the current graph construction each node has a one-dimensional feature (gene expression value), the SHAP-like score can be interpreted directly as a patient-specific gene importance score. To obtain more robust

class-level summaries, these node scores are averaged across multiple graphs belonging to the same class. The final output consists of the top-ranked genes for each class.

Expected output. The SHAP-like explainer is expected to identify genes whose patient-specific expression values contribute most strongly to the classification decision. These outputs are therefore particularly useful for ranking candidate biomarkers and for comparing model-derived importance with external biological evidence such as differential expression signatures. Architecture is shown in Fig. 4.9

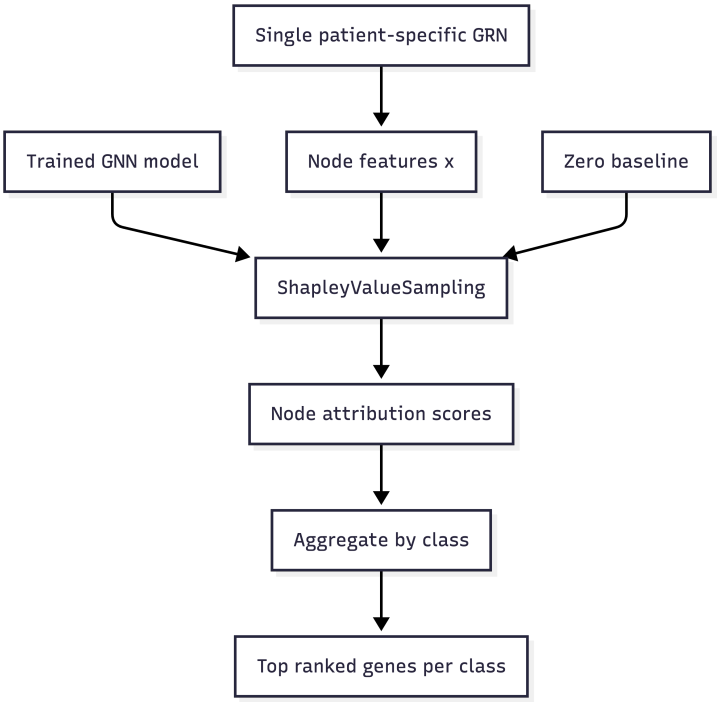


Figure 4.9: SHAP architechure.

4.9.2 GNNExplainer

GNNExplainer is used to obtain structure-aware explanations of graph-level predictions. Unlike the SHAP-like approach, which focuses on feature attribution, GNNExplainer learns soft masks over nodes and edges in order to identify a compact explanatory subgraph together with the most relevant node attributes.

In the implemented pipeline, GNNExplainer is configured at the **graph**

classification level, consistently with the downstream prediction task. The explainer is applied to one patient-specific GRN at a time and receives as input:

- node features x,
- graph connectivity `edge_index`,
- graph-level batch vector,
- regulatory edge weights `edge_weight`.

The explanation procedure produces:

- a **node mask**, quantifying the importance of nodes and node attributes;
- an **edge mask**, quantifying the importance of regulatory interactions for the prediction.

In the current implementation, node-level importance is aggregated across graphs of the same class by averaging the absolute node-mask scores. In addition, graph-specific masks are saved to disk for downstream inspection. To support biological interpretation at pathway level, the averaged node-importance scores are further mapped to curated gene sets loaded from a GMT file. For each pathway, a pathway score is obtained by summing the importance scores of the genes belonging to that pathway. This produces a class-specific ranking of relevant pathways.

Expected output. GNNExplainer is expected to reveal not only important genes, but also the *regulatory context* in which they act. The resulting explanations can therefore highlight small regulatory modules, influential edges, and pathway-level patterns that are not directly accessible through feature-only attribution methods. Architecture is shown in Fig. 4.10

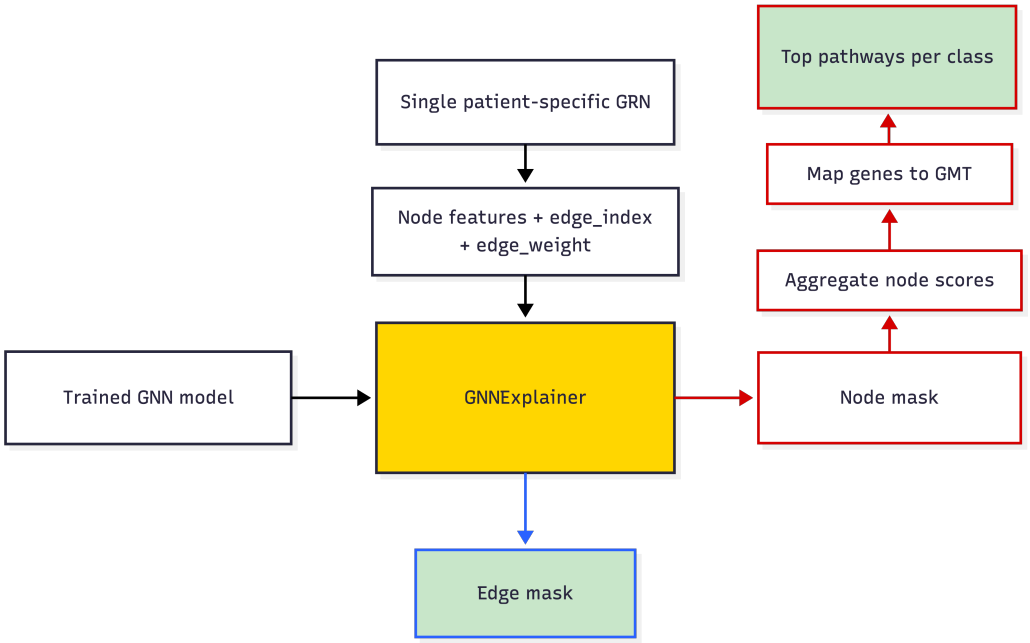


Figure 4.10: GNN Explainer architechure. Colored borders indicate the processes to reach the two outputs: edge mask and top pathways.

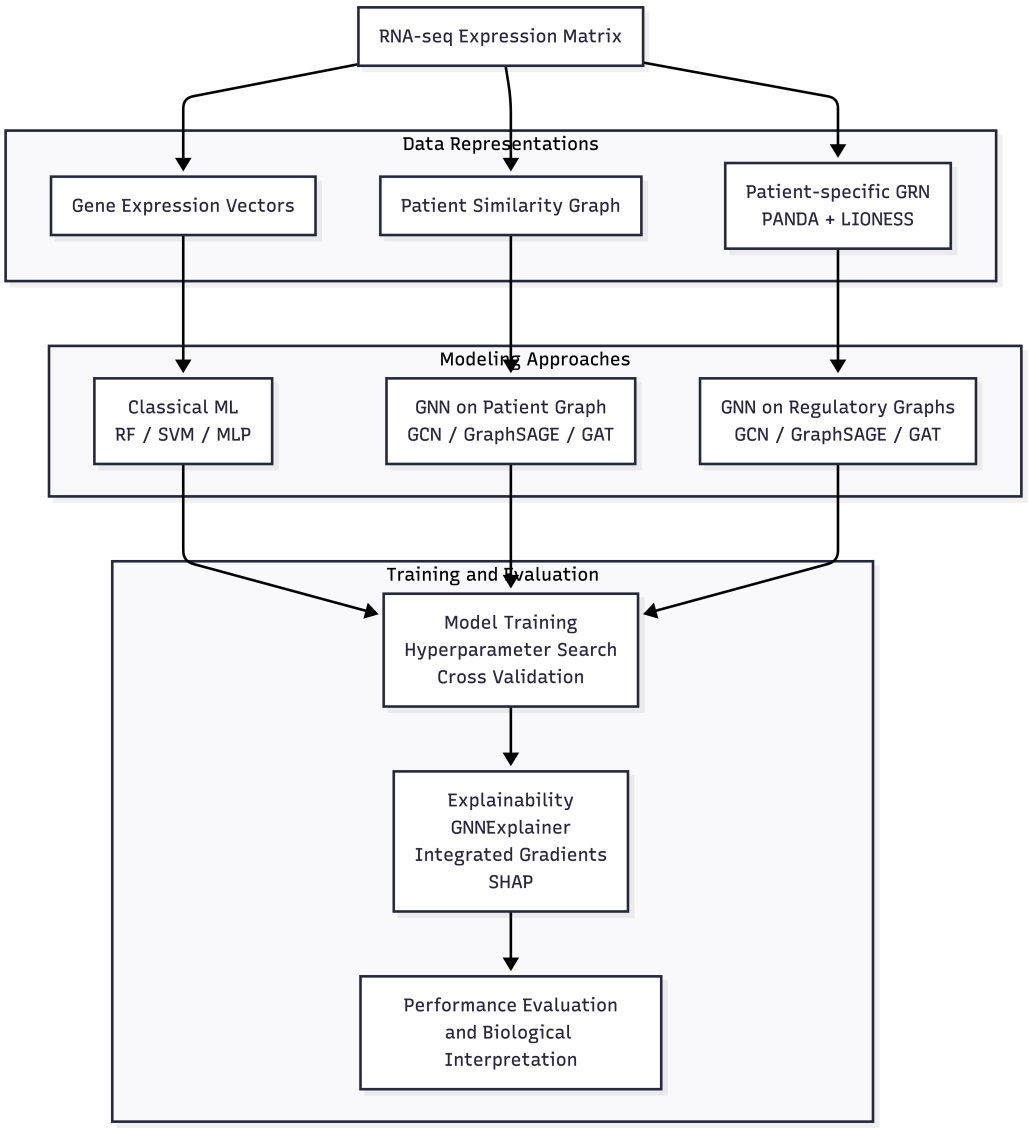


Figure 4.1: Chapter 4 modeling framework. RNAseq expression is transformed into three representations: (A) gene expression vectors for classical ML baselines, (B) patient–patient similarity graphs for GNN-based sample modeling, and (C) patient-specific LIONESS GRNs for regulatory-graph learning. Cross-cancer experiments require cohort size reduction via centrality- or diversity-based sampling.

Chapter 5

Results and Discussion

This chapter presents the empirical evaluation of the proposed modeling pipelines for leukemia subtype classification and cross-cancer disease association analysis. Results are organized according to the three transcriptomic data representations introduced in Chapter 4.2:

- **Representation A:** Gene expression vectors analyzed using classical ML and DL models. Section 4.2.1
- **Representation B:** Patient–patient similarity graphs analyzed using GNNs. Section 4.2.2
- **Representation C:** Patient-specific GRNs inferred with PANDA and LIONESS and analyzed using GNNs. Section 4.2.3

This progressive organization reflects the increasing structural complexity of the data representation, moving from feature-centric transcriptomic models toward regulatory network modeling. The final sections of this chapter further investigate model explainability and cross-cancer generalization using breast cancer transcriptomic data.

5.1 Overview of Experimental Evaluation

5.1.1 Prediction Tasks

Two prediction settings are considered throughout the experiments:

- **Binary classification**: Tumor vs Normal samples
- **Multiclass classification**: Normal, Acute Lymphoblastic Leukemia (ALL), and Acute Myeloid Leukemia (AML)

The multiclass setting reflects the clinical heterogeneity of leukemia subtypes, which exhibit distinct transcriptional programs and regulatory alterations.

5.1.2 Dataset Composition

The primary experiments are conducted on the leukemia cohort described in Chapter 3. The dataset contains RNA sequencing profiles retrieved from the Genomic Data Commons repository.

Table 5.1: Leukemia cohort composition used in the experiments

Class	Number of samples	Description
Normal	380	Healthy control samples
ALL	349	Acute Lymphoblastic Leukemia
AML	1645	Acute Myeloid Leukemia
Total	2374	RNAseq samples

Due to the imbalance between leukemia subtypes, evaluation metrics that account for class distribution are used.

Evaluation metrics. Above those investigated: accuracy, Macro f1-score and **balanced accuracy**, the latter is considered the primary evaluation metric for model comparison, as the AML subtype represents the majority of samples in the dataset.

5.2 RNAseq Data Exploration and Feature Representation

Before training predictive models, an exploratory analysis of the RNA sequencing dataset was conducted to assess the global structure of the transcriptomic data and to evaluate the effect of normalization strategies on sample separability.

RNAseq data are characterized by high dimensionality and strong differences in expression scale across genes, which can affect both visualization and downstream modeling [16, 46]. To assess the effect of preprocessing on the global structure of the dataset, dimensionality reduction was applied under different normalization strategies.

Three input representations were evaluated:

- Raw FPKM expression values

- Log-transformed expression values $\log_2(\text{FPKM} + 1)$
- Log-transformed values followed by z-score normalization

The effect of the log transformation on the overall expression distribution is shown in Figure 5.1. The corresponding low-dimensional projections are reported in Figure 5.2.

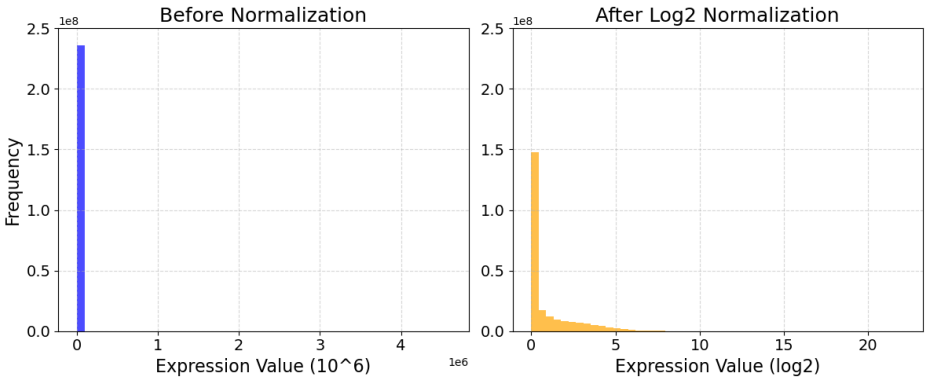


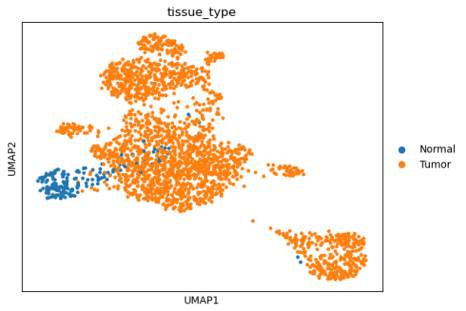
Figure 5.1: Distribution of RNAseq expression values before and after logarithmic transformation. The log transform compresses the dynamic range and reduces the dominance of highly expressed genes.

Raw FPKM values show limited class separation, with the projection largely influenced by a small number of highly expressed genes. The $\log_2(\text{FPKM}+1)$ transformation improves the visual organization of the samples by stabilizing the expression range and reducing skewness. Z-score normalization produces a comparable global structure, but may attenuate absolute expression differences that remain biologically informative for classification. Numerical results are shown in Table. 3.4

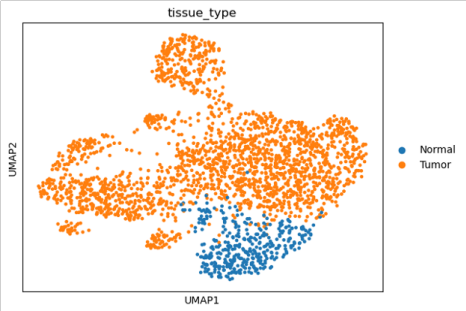
Based on these observations, $\log_2(\text{FPKM}+1)$ **was selected as the main input representation** for downstream modeling.

5.3 Gene Expression-Based Models

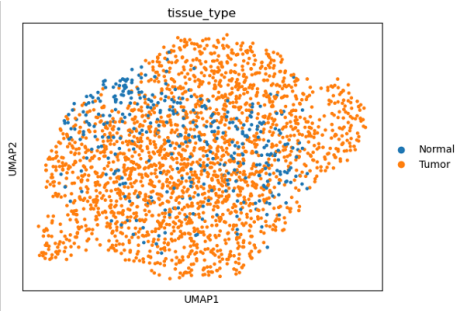
Gene expression-based models represent the most direct approach for transcriptomic classification tasks. In this representation, each gene is treated as an independent feature and the RNAseq matrix is modeled as a high-dimensional tabular dataset. This paradigm has been widely adopted in cancer transcriptomics and has demonstrated strong predictive performance for tumor subtype classification [31, 79].



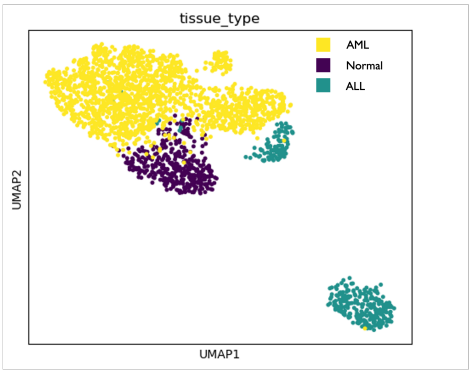
(a) FPKM



(b) $\log_2(\text{FPKM} + 1)$



(c) $\log_2(\text{FPKM}+1)$ with z-score normalization



(d) $\log_2(\text{FPKM}+1)$ - MULTICLASS

Figure 5.2: Low-dimensional projections of the RNAseq expression matrix under different normalization strategies. Each point represents one sample, colored by class label. Last image (d) corresponds to Multiclass problem

In this study, gene expression models serve as a baseline for evaluating the additional contribution of graph-based representations introduced in later sections.

5.3.1 Dimensionality Reduction

The original RNAseq matrix contains almost 60,000 genes, resulting in a very high-dimensional feature space. To reduce dimensionality while preserving the majority of transcriptomic variance, Principal Component Analysis (PCA) was applied to the log-transformed expression matrix.

The first 1116 principal components were retained, capturing the 90% of global variance across the dataset.

This reduced representation was used as input for classical ML and neural network models.

Table 5.2: Gene expression feature representation after PCA dimensionality reduction

Dataset property	Value
Number of samples	2374
Original number of genes	∼60,000
PCA components retained	1116
Final input matrix	2374×1116

5.3.2 Binary Classification Results

The first prediction task consists of distinguishing tumor samples from normal controls. Several classical ML algorithms were evaluated, including RF, SVMs, and neural network models. Best model configurations, resulting from grid search can be found on the project repository [29]. Results are shown in Table 5.3

Table 5.3: **Binary classification** results using gene expression features

Model	Accuracy	Bal. Accuracy	Macro F1
RF	0.9	0.89	0.94
SVM	1	**1**	1
MLP Classifier	1	0.99	0.99
Feedforward Neural Network	1	**1**	0.99

Gene expression models achieve strong predictive performance in the binary setting, reflecting the large transcriptional differences between healthy and leukemia samples.

5.3.3 Multiclass Leukemia Subtype Classification

The second task consists of multiclass classification across the three biological classes:

- Normal
- Acute Lymphoblastic Leukemia (ALL)
- Acute Myeloid Leukemia (AML)

Compared with the binary setting, the multiclass problem is generally more challenging due to partial overlap between leukemia transcriptional programs.

Table 5.4: **Multiclass** leukemia subtype classification using gene expression features

Model	Accuracy	Bal. Accuracy	Macro F1
RF	0.95	0.79	0.85
SVM	0.99	0.99	0.99
MLP Classifier	0.98	0.98	0.98
Feedforward Neural Network	1	**1**	1

Despite the increased complexity of the task, expression-based models maintain exceptional performance results, remaining capable of capturing discriminative transcriptomic signatures associated with leukemia subtypes. FFN shown maximum *balanced score* for both binary and multiclass problems.

5.3.4 Discussion

The results confirm that gene expression alone provides strong predictive signals for leukemia classification. This observation is consistent with previous studies demonstrating that transcriptomic profiles encode disease-specific transcriptional programs and can be used to accurately distinguish cancer subtypes [31, 54].

To further investigate the biological relevance of the expression-based baseline, the most important genes identified by the RF model were examined. Figure 5.3 reports the top 20 ranked genes contributing to the classification model.

A qualitative inspection of these genes reveals a mixed pattern. Some genes have previously been associated with leukemia biology or related oncogenic processes. For example, *BIRC6* has been linked to anti-apoptotic mechanisms and adverse clinical features in acute leukemia, while *CAMK2G* has been implicated in the regulation of leukemia stem-like cell survival in acute myeloid leukemia [18, 45]. Similarly, *SIX6* has been reported as a transcription factor involved in regulatory programs in T-cell acute lymphoblastic leukemia [51], and *MCOLN2* has been described among genes with potential functional relevance in leukemia transcriptomic studies [9].

At the same time, a large fraction of the top-ranked (16 out of 20) features corresponds to poorly characterized loci, pseudogene-like annotations, or genes without strong leukemia-specific evidence in the literature. This suggests that the expression-based classifier captures a mixture of biologically meaningful signals and broader transcriptomic variability.

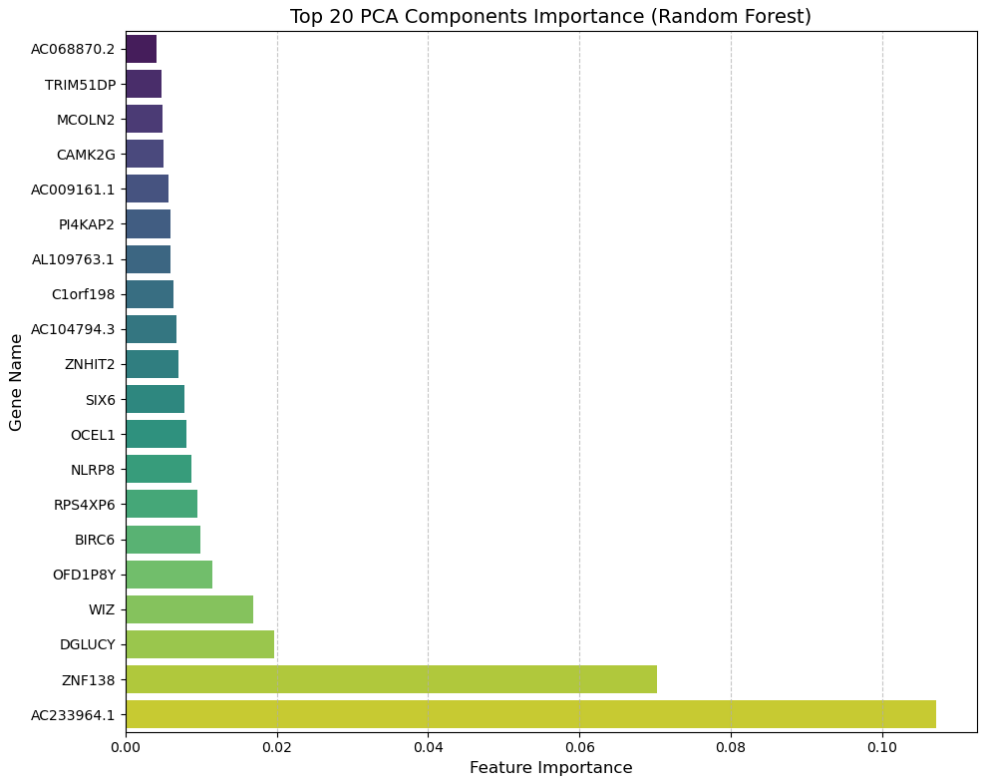


Figure 5.3: Top 20 genes identified by the RF model using PCA-based gene expression features. Gene importance values correspond to the contribution of each feature to the classification decision.

Importantly, expression-based models treat genes as independent predictors and do not explicitly capture regulatory interactions between transcription factors and their targets. Consequently, while these models achieve high predictive performance, they provide limited insight into the regulatory mechanisms underlying disease heterogeneity.

These observations support the hypothesis that regulatory network representations may provide improved biological interpretability by explicitly modeling gene-gene regulatory interactions. This motivates the exploration of graph-based and GRN-based representations introduced in the following sections.

5.4 Patient–Patient Similarity Graph Models

In addition to expression-based classifiers, a graph representation of the cohort was evaluated in order to incorporate relationships between samples.

Table 5.5: Selected genes from the RF feature importance analysis with literature evidence related to leukemia biology.

Gene	Reported biological role	Reference
BIRC6	Anti-apoptotic regulator associated with leukemia progression	[45]
CAMK2G	Signaling regulator involved in AML stem cell survival	[18]
SIX6	Transcription factor implicated in T-ALL regulatory networks	[51]
MCOLN2	Gene reported in leukemia transcriptomic analyses	[9]

In this setting, each patient sample is represented as a node in a graph, while edges encode transcriptomic similarity between samples as described in Section 4.2.2.

Graph representations allow learning algorithms to exploit relational structure within the cohort, potentially capturing patterns shared among patients with similar transcriptional programs.

The patient similarity graph constructed from the leukemia cohort contains the following structural properties:

Table 5.6: Patient similarity graph statistics

Graph property	Value
Number of nodes	2374
Number of edges	2000
Graph type	Undirected
Node features	PCA gene expression components

GNNs were trained on this representation to evaluate whether relational information between samples improves classification performance.

To qualitatively inspect the structure learned by the graph model, the node embeddings generated by the GNN were projected into two dimensions using UMAP, as shown in Figure 5.4. Each point corresponds to a patient sample colored by its class label.

Overall, the performance of patient graph models remains comparable to that obtained using direct expression-based classifiers - Table 5.7. This suggests that most of the discriminative signal is already contained in the individual gene expression profiles, while relational information between patients provides limited additional benefit.

Table 5.7: Performance of GNN models on the patient similarity graph for binary and multiclass classification.

	Binary			Multiclass		
Model	Acc.	Bal-Acc.	Macro F1	Acc.	Bal-Acc.	Macro F1
GCN	0.93	0.93	0.96	0.86	0.85	0.89
GraphSAGE	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
GAT	0.92	0.92	0.96	0.93	0.47	0.63

One possible explanation is that patient similarity graphs mainly capture global transcriptomic similarity rather than regulatory mechanisms underlying disease heterogeneity. Consequently, while this representation introduces relational structure, it may not fully reflect the biological interactions driving leukemia progression.

These observations motivate the exploration of regulatory network-based representations, where edges correspond to transcription factor-gene regulatory interactions rather than sample similarity. Such representations may provide a more biologically meaningful graph structure for learning disease-related patterns.

5.5 Global Gene Regulatory Network Reconstruction

GRNs provide a systems-level representation of transcriptional regulation by modeling interactions between transcription factors and their target genes. Unlike expression-based representations that treat genes as independent variables, GRNs explicitly encode regulatory dependencies between genes and transcription factors.

In this work, GRNs were reconstructed using the PANDA algorithm, which integrates three complementary sources of biological information: transcription factor binding motifs, protein–protein interaction data between transcription factors, and gene expression profiles [33]. The resulting network represents inferred regulatory interactions between transcription factors and target genes across the cohort.

The global PANDA network serves as the reference regulatory structure used for subsequent reconstruction of patient-specific networks through the LIONESS framework.

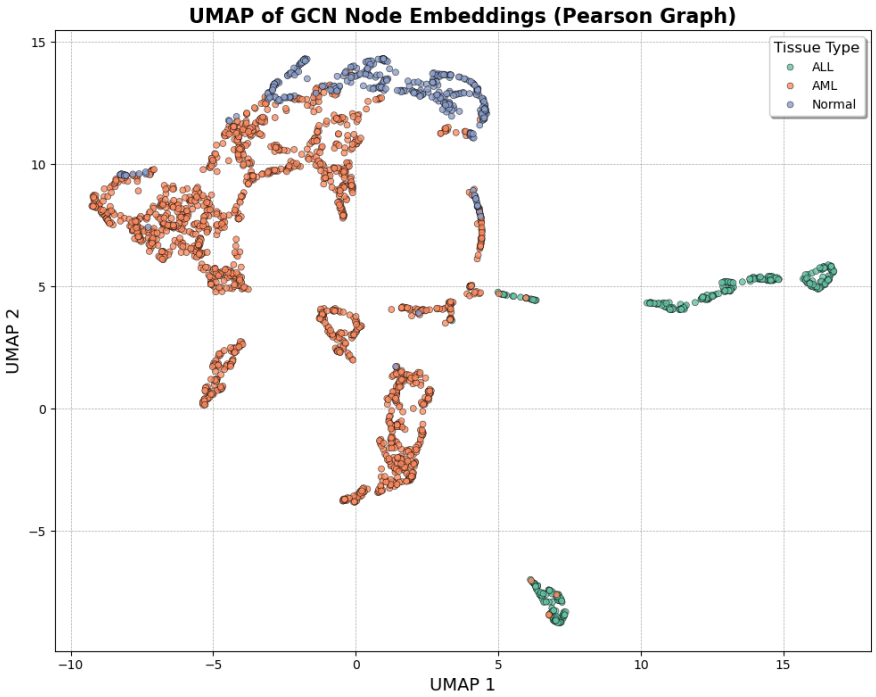


Figure 5.4: UMAP projection of patient embeddings learned from the similarity graph. Each point represents a sample colored by class label.

5.5.1 Network Structure

The PANDA algorithm produces a weighted bipartite network linking transcription factors to their predicted target genes. Edges represent inferred regulatory influence scores, commonly referred to as *regulatory force* values.

Table 5.8: Global PANDA regulatory network statistics. Knowledge-like strategy. *Density is calculated as $interactions/set(genes * tf)$

Network property	Tumor GRN	Normal GRN
Number of transcription factors	2,590	1,821
Number of target genes	1,758	1,640
Total regulatory interactions	4,552,320	2,985,585
Network density*	0.383	0.440
Edge weight type	Regulatory force score	Regulatory force score

The resulting network represents a dense regulatory structure capturing

potential transcriptional regulation across the leukemia cohort. Tumor and Normal GRNs statistics are visible in Table 5.8

5.5.2 Distribution of Regulatory Interaction Strength

The distribution of PANDA regulatory force scores provides insight into the global structure of the inferred regulatory network.

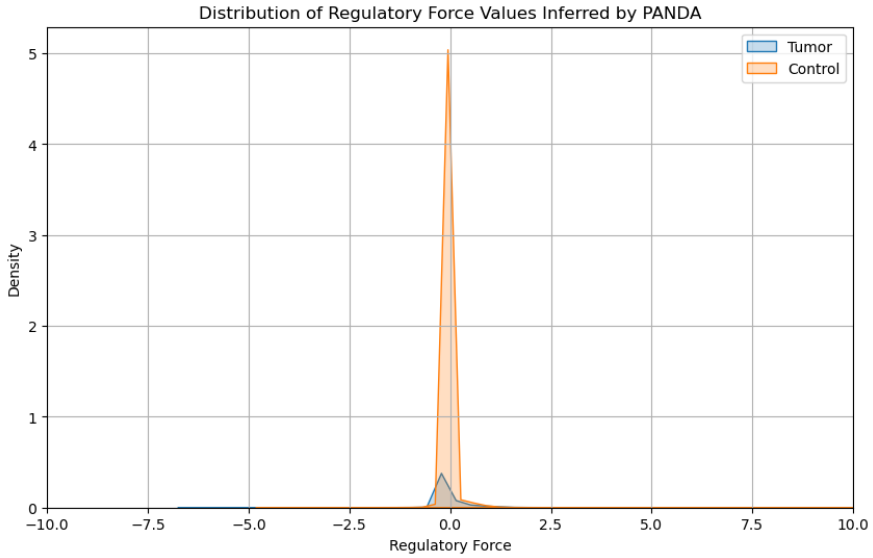


Figure 5.5: Distribution of regulatory force values inferred by PANDA across all transcription factor–gene interactions.

Most interactions cluster around low absolute force values, while a smaller subset of edges exhibits stronger regulatory influence. This pattern is consistent with previous observations in reconstructed GRNs, where only a fraction of transcription factor–gene relationships correspond to strong regulatory interactions.

5.5.3 Hub Transcription Factors

To identify key regulators within the reconstructed network, transcription factors were ranked according to their regulatory influence across target genes. Fig. 5.6 Hub transcription factors correspond to regulators with high centrality or cumulative regulatory influence within the network.

Several high-ranking transcription factors correspond to regulators previously implicated in hematopoietic differentiation and leukemia biology. Such

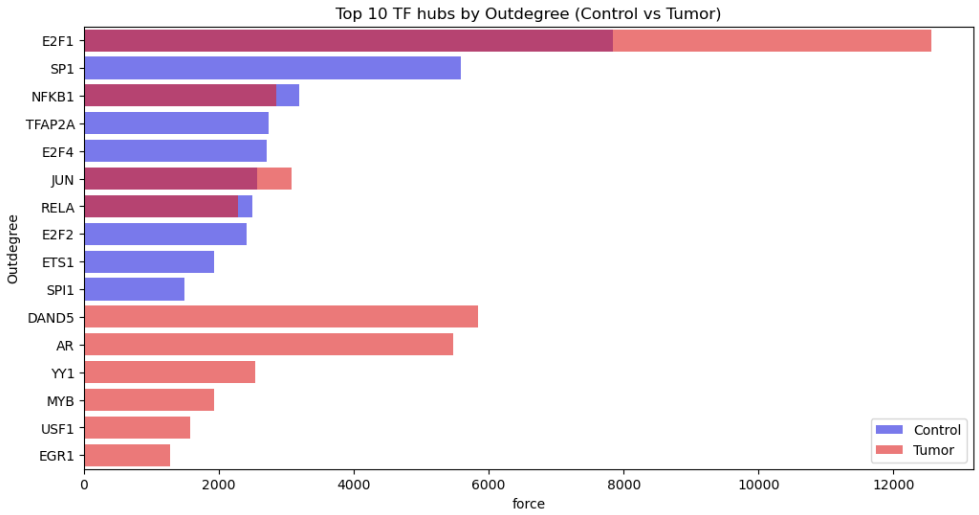


Figure 5.6: Top transcription factors ranked by regulatory influence in the PANDA network.

regulators are known to play central roles in transcriptional programs governing blood cell development and malignant transformation [11].

5.5.4 Highly Regulated Target Genes

In addition to transcription factor hubs, genes receiving strong cumulative regulatory influence were also examined. These genes represent highly regulated targets potentially involved in disease-related transcriptional programs. Fig. 5.7

Highly regulated genes may correspond to downstream effectors of transcriptional programs associated with leukemia pathogenesis.

5.5.5 Biological Interpretation

The reconstructed PANDA networks provide a biologically more informative view of leukemia transcriptomes than direct expression-based feature rankings. Top 10 TF hubs and highly regulated genes are shown in Fig. 5.6 - 5.7, respectively.

In the tumor GRN, several top hub transcription factors are strongly consistent with known leukemia biology, including *E2F1*, *JUN*, *NFKB1*, *RELA*, and *SPI1*. *E2F1* has been reported as aberrantly upregulated in AML and linked to proliferative transcriptional programs, whereas *JUN* is frequently overexpressed in AML and supports leukemic cell survival. Likewise, the NF-

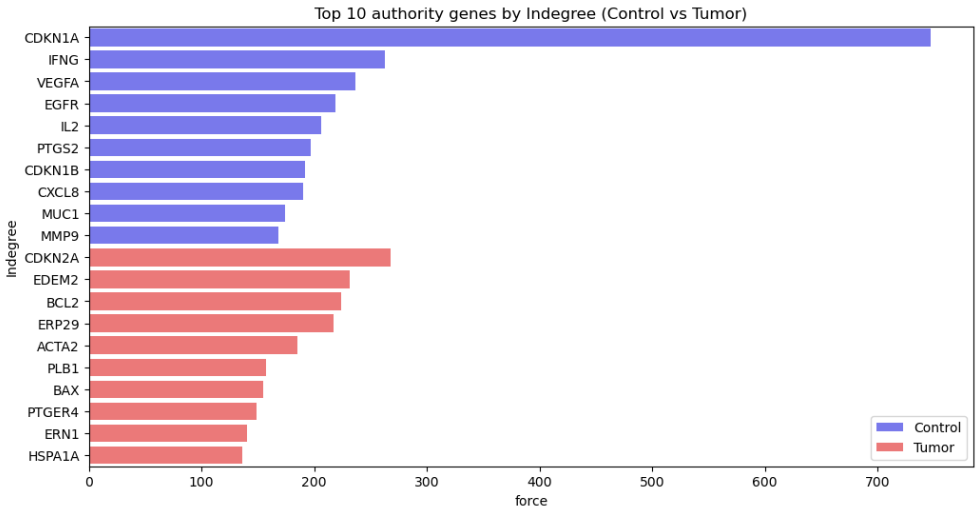


Figure 5.7: Top target genes receiving the highest cumulative regulatory influence in the PANDA network.

κB pathway, represented here by *NFKB1* and *RELA*, is widely implicated in AML pathogenesis and has also been associated with ALL biology. *SPI1* (PU.1) is a core regulator of hematopoietic differentiation whose deregulation is classically linked to leukemogenesis [15, 27, 71, 75, 98, 100].

By contrast, some other highly ranked factors, such as *TFAP2A*, *E2F4*, *DAND5*, *AR*, and partly *USF1*, appear less specifically supported in leukemia and are more likely to reflect general transcriptional activity or dataset-specific effects. Nevertheless, the overall pattern remains coherent: compared with the normal GRN, the tumor network is enriched in hubs associated with proliferation, stress adaptation, and survival signaling, while the normal network retains regulators more compatible with physiological hematopoietic or homeostatic programs, including *EGR1* and *MYB* [26, 49].

A similarly informative contrast emerges from the highly regulated target genes. In the tumor GRN, top authority genes include *CDKN2A*, *BCL2*, and *BAX*, pointing to dysregulation of cell-cycle control and apoptosis, both central processes in leukemia. The presence of *BCL2* is particularly relevant, given its established importance in leukemic cell survival and therapeutic targeting. In addition, genes such as *ERN1*, *ERP29*, *EDEM2*, and *HSPA1A* suggest activation of endoplasmic reticulum stress and unfolded protein response-related programs, which is compatible with stress-adaptive states in malignant cells [61, 100].

In the normal GRN, the most regulated genes are instead dominated by *CDKN1A*, *IFNG*, *VEGFA*, *IL2*, *PTGS2*, *CXCL8*, and *MMP9*. Taken to-

gether, these genes point more toward immune signaling, cytokine response, and microenvironment-related processes than toward leukemia-specific oncogenic control. Therefore, the contrast between the two PANDA networks is biologically plausible: the tumor network emphasizes survival and transformation-related programs, whereas the normal network retains a more immune-regulatory and homeostatic signature [32, 72].

Overall, these findings support the central hypothesis of this work. While expression-based models are highly predictive, their top-ranked features only partially overlapped with well-established leukemia biology. By contrast, the PANDA GRN highlights regulators and target genes that align more directly with known mechanisms of leukemogenesis, including cell-cycle dysregulation, NF-κB signaling, hematopoietic transcriptional control, apoptosis, and stress adaptation. This indicates that GRN representations may provide improved biological interpretability even when predictive performance alone does not necessarily exceed that of direct gene expression models. The global PANDA network therefore provides a biologically grounded basis for the subsequent reconstruction of patient-specific regulatory networks using LIONESS [11, 48].

5.6 Patient-Specific Networks with LIONESS

The LIONESS procedure generated one weighted regulatory network per sample. All networks share the same node space, defined by the transcription factors and target genes included in the PANDA prior, while edge weights vary across patients according to the inferred sample-specific regulatory contribution.

Because each LIONESS network contains a very large number of TF–gene interactions, direct visualization of full patient-specific graphs is not informative. Therefore, network-level summary statistics and regulator-level aggregated scores were used to characterize inter-patient variability in a compact and interpretable manner. Fig. 5.8

5.7 GNNs on Patient-Specific GRNs

Patient-specific regulatory networks reconstructed with LIONESS were used as graph inputs for GNN models. In this representation, each patient is described by an individual transcription factor–GRN derived from the global PANDA model.

Each graph contains the same set of nodes (transcription factors and

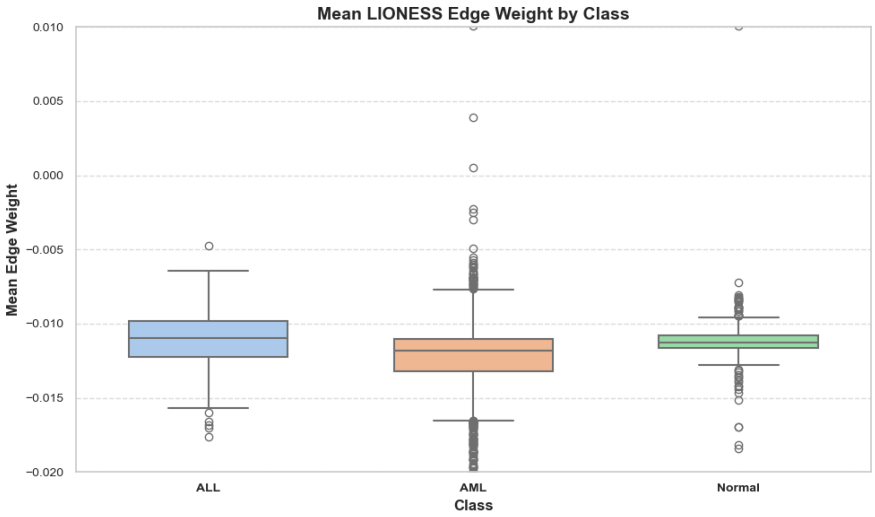


Figure 5.8: Boxplot of mean edge weights across patient-specific LIONESS networks.

target genes), while edge weights correspond to sample-specific regulatory force values inferred by LIONESS [48]. Node features were derived from normalized gene expression values associated with each gene node.

5.7.1 Graph Dataset Overview

The resulting dataset consists of a collection of patient-specific graphs sharing a common topology but differing in edge weights. These graphs constitute the input for graph classification models. Statistics can be seen at table 5.9.

Table 5.9: Summary statistics of the patient-specific GRN dataset used for graph classification.

Property	Knowledge-like	Sequence -like
Number of graphs (patients)	2374	2374
Number of unique genes per graph	4797	3093
Number of unique tfs per graph	5837	654
Number of nodes per graph	10,634	3747
Number of edges per graph	28,000,089	2,022,822
Node feature dimension	1 (gene expression)	1 (gene expression)
Graph type	Directed TF–gene	Directed TF–gene
Edge weights	filter force >0	filter force >0

5.7.2 Model Selection Workflow

The final classifier was selected through a progressive evaluation pipeline:

1. Evaluation of knowledge and Sequence PANDA GRNs using five different graphs. Model used : GCN
2. Baseline GNN comparison,
3. Evaluation of advanced GNN architectures:,
4. election of the final model based on predictive performance.

Steps are resumed in the Fig. 5.9

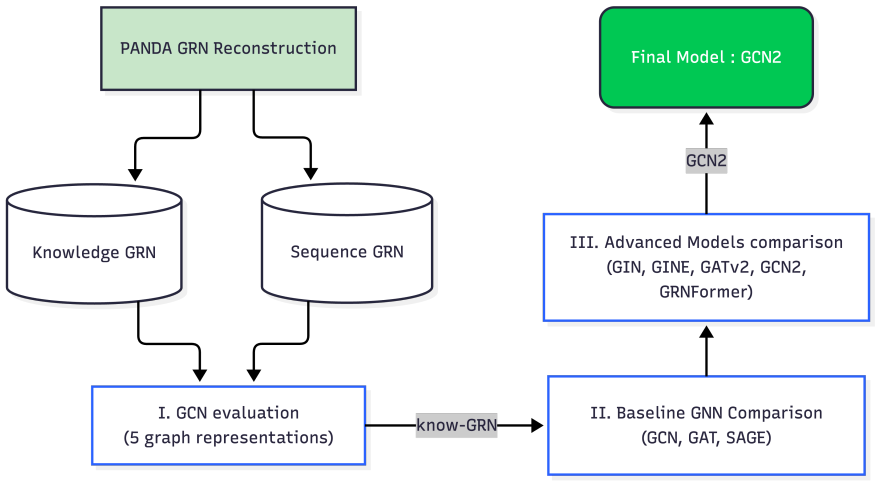


Figure 5.9: Experimental model selection pipeline leading to the final GCN2 classifier.

5.7.3 Evaluation of PANDA Network Construction Strategies

Before training GNNs, different PANDA network construction strategies were evaluated for the *binary case*. In particular, two regulatory priors were considered: a sequence-based prior and a knowledge-based prior. Additionally, multiple graph construction settings were tested as summarized in Table 4.5.

To determine the most suitable configuration, a baseline Graph Convolutional Network (GCN) classifier was trained on each resulting graph representation.

Table 5.10: Evaluation of graph construction strategies using GCN on patient-specific GRNs in binary classification. Balanced accuracy is selected as main metrics

Test configuration	PANDA Sequence	PANDA Knowledge
Test 0	0.5	0.5
Test 1	0.5	0.5
Test 2	0.51	0.599
Test 3	0.5	**0.717**
Test 4	0.49	0.716

Since problem is binary a balanced score of 0.5 means that model is classifying all data to 1 class. The knowledge-based PANDA prior consistently achieved higher balanced accuracy especially in the last 3 configurations. In addition to improved predictive performance, the knowledge-based network provides a biologically grounded representation derived from curated regulatory information. For these reasons, this configuration was selected for all subsequent experiments.

5.7.4 Comparison of GNN Architectures

Using the selected PANDA knowledge-based regulatory network representation, multiple GNN architectures were evaluated for graph classification.

Table 5.11: Performance comparison of **baseline** GNN architectures on patient-specific GRNs. PANDA-Knowledge strategy

	Binary Classification			Multiclass Classification		
Model	Test 2	Test 3	Test 4	Test 2	Test 3	Test 4
GCN	0.59	**0.72**	0.72	0.51	**0.60**	0.55
GraphSAGE	0.58	0.5	0.5	0.5	0.57	0.57
GAT	0.5	0.5	0.5	0.33	0.33	0.33

In the binary classification task, only the GCN architecture achieved good performance. However, the multiclass problem proved significantly more challenging, reflecting the heterogeneity between leukemia subtypes, Tab.5.11. In this classification there are 3 classes. A balanced score = 0.33 means the model is classifying all samples to 1 class.

Other models were tested in the PANDA sequence strategy for training

time: 2 h per test vs 18.4 h per test for the knowledge strategy. Multiclass problem is shown in Tab.5.12

Table 5.12: Comparison of **advanced** GNN models. Multiclass problem on PANDA sequence strategy.

Model	Test 2	Test 3
GCN2	**0.48**	**0.46**
GAT2	0.40	0.46
GIN	0.33	0.33
GINE	0.34	0.34
GRN Former	0.35	0.35

Among the tested models, GCN2 provided the most stable performance also for knowledge startegy and was therefore selected for further hyperparameter optimization.

5.7.5 Hyperparameter Optimization

Hyperparameter optimization was performed to identify the best-performing configuration of the GCN2 model on the patient-specific GRN dataset. Instead of reporting all explored configurations, Table 5.13 summarizes the final parameter set achieving the highest validation balanced accuracy.

The selected configuration reflects the high class imbalance of the dataset. In particular, focal loss combined with inverse-frequency class weighting was found to improve model stability and balanced accuracy during training. The optimal validation performance was reached at epoch 20, after which the model began to show signs of performance saturation.

5.7.6 Final Cross-Validation Results

The final model configuration was evaluated using cross-validation in order to obtain a robust estimate of generalization performance.

Table 5.13: Best hyperparameter configuration obtained during model selection for the GCN2 architecture.

Parameter	Value
Epochs	40
Hidden dimension	128
Dropout	0.2
Learning rate	5×10^{-4}
Weight decay	0.0
Optimizer	AdamW
Class weighting	Inverse frequency
Effective β	0.99
Loss function	Focal loss
Label smoothing	0.05
Focal γ	2.0
Gradient clipping	0.0
Edge weight processing	$\text{clip}_{[0,5]}$
Best validation balanced accuracy	0.719
Best validation macro F1-score	0.591
Best epoch	20
Training time	7.7 hours

Table 5.14: Final cross-validation results of the selected GCN2 model on patient-specific GRNs (multiclass classification).

Fold	Accuracy	Precision	Recall	Macro F1	Balanced Acc.
1	0.644	0.589	0.671	0.606	0.671
2	0.556	0.560	0.689	0.550	0.689
3	0.589	0.566	0.698	0.572	0.698
4	0.686	0.615	0.718	0.642	0.718
5	0.582	0.565	0.702	0.568	0.702
Average	**0.611**	**0.579**	**0.696**	**0.588**	**0.696**

The final GCN2 model achieved an average balanced accuracy of 0.696 and a macro F1-score of 0.588 in the multiclass classification task. Bal-

anced accuracy remained relatively stable across folds, ranging from 0.671 to 0.718, indicating consistent model performance despite the heterogeneity of leukemia subtypes.

The higher recall compared to precision suggests that the model captures most disease-related regulatory patterns while occasionally confusing closely related leukemia classes.

The higher recall compared to precision suggests that the model captures most disease-related regulatory patterns while occasionally confusing closely related leukemia subtypes.

Model checkpoints and full training configurations were saved for the best-performing runs. These artifacts were subsequently used as the input models for the explainability analyses presented in the explainability section.

5.7.7 Graph Embedding Visualization

To qualitatively inspect the learned representations, graph embeddings from the final GCN2 model were projected to two dimensions using UMAP.

Compared with previous representations, the embedding space reveals partial separation between disease classes while still exhibiting substantial overlap between leukemia subtypes, reflecting the intrinsic biological complexity of the multiclass task.

5.7.8 Leukemia Classifier Discussion

The final GCN2 model achieved an average balanced accuracy of 0.696 and a macro F1-score of 0.588 in the multiclass classification task. Although these values may appear moderate compared with some results reported in the literature, the present problem is considerably more challenging than many commonly studied transcriptomic classification settings. In particular, the model must discriminate among Normal, ALL, and AML samples using patient-specific GNNs inferred from RNAseq data rather than directly optimized tabular gene expression features. Several studies reporting higher performance operate in substantially different experimental conditions. For example, Ramirez et al. obtained 94.6% accuracy using graph convolutional networks on TCGA data for cancer-type prediction across multiple solid tumors, a task characterized by strong tissue-of-origin transcriptional signals [73]. Similarly, the HallmarkGraph framework reported cross-validation accuracies between 85% and 99% in a large pan-cancer cohort combining transcriptomic data with curated hallmark-informed regulatory structure [97]. These settings benefit from stronger inter-class separation

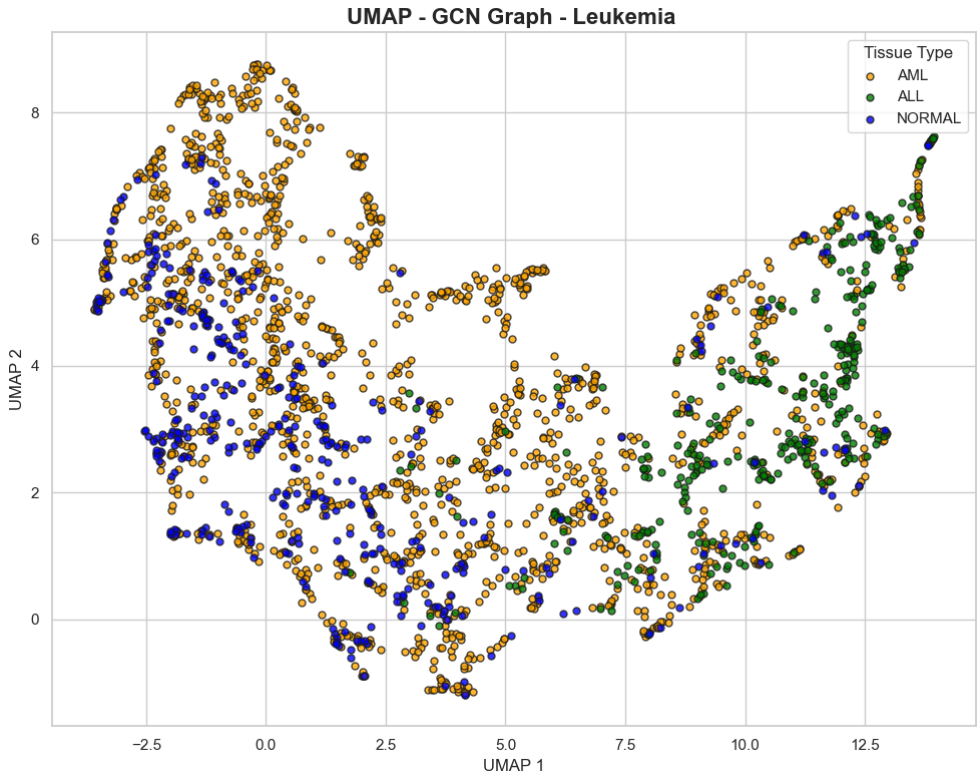


Figure 5.10: UMAP projection of graph embeddings learned from patient-specific GRNs.

and are therefore not directly comparable to a multiclass leukemia subtype prediction problem.

In the present work, each patient is represented as a LIONESS-derived TF-gene GRN, where node features correspond to gene expression and edge weights represent sample-specific regulatory force values. This representation preserves biologically meaningful regulatory structure but introduces an additional inference layer between raw transcriptomic measurements and the classification model. Moreover, leukemia subtypes share broad transcriptional programs associated with hematopoietic proliferation and survival, leading to substantial overlap between classes, as also reflected by the partial separation observed in the UMAP projection of graph embeddings, Fig. 5.10. Similar challenges have been highlighted in recent reviews of GNNs applied to cancer genomics, where differences in graph construction, omics modalities, and classification tasks make direct numerical comparisons difficult [34, 105]. Other studies combining multi-omics data and sample similarity graphs also report strong performance for molecular subtype classification, but rely on

richer feature spaces and different graph representations than the patient-specific GRN framework used here [93]. Taken together, these considerations suggest that the observed performance reflects the inherent complexity of regulatory-network-based leukemia classification rather than limitations of the modeling approach itself.

5.8 Cross-Cancer Classification with Patient-Specific GRNs

After establishing the graph classification pipeline on the leukemia cohort, the same modeling framework was extended to the cross-cancer setting. In this experiment, the selected GCN2 architecture was applied without structural modifications to a joint cohort including leukemia, breast cancer, and normal samples. The objective was to evaluate whether the regulatory-network-based representation could generalize beyond leukemia subtype classification and support a broader disease-association setting.

Compared with the leukemia-only task, the cross-cancer dataset introduces greater biological variability, as samples originate from distinct tissues and malignancies. At the same time, the classification objective is simplified from subtype discrimination to a broader disease-association paradigm, where the model is expected to separate tumor from normal states across heterogeneous tissue contexts.

Patient-specific regulatory networks were reconstructed using the same PANDA–LIONESS pipeline adopted for the leukemia cohort, preserving a consistent graph generation strategy across experiments.

5.8.1 Cross-Cancer GRN Dataset Overview

The final cross-cancer GRN dataset is summarized in Table 5.15. All patient-specific graphs share a common regulatory structure, while edge weights vary according to the sample-specific LIONESS estimates.

Cross-Cancer Dataset Construction

Since the leukemia cohort is substantially larger than the breast cancer dataset, a subsampling strategy was applied. Specifically, the centrality-based procedure (section 4.3) was used to select a representative subset of leukemia samples from the original dataset. This approach retains samples that occupy central positions in the expression space, ensuring that the re-

Table 5.15: Summary statistics of the cross-cancer patient-specific GRN dataset.

Property	Value
Number of graphs (patients)	2102
Number of unique genes	5836
Number of unique tf	7877
Number of nodes	13,713
Number of edges	45,970,172
Node feature dimension	1
Graph type	Directed TF–GRN
Edge weights	filter force >0

duced cohort preserves the main transcriptional structure of the leukemia population.

Both tumor and normal samples were retained for each cancer type.

Table 5.16 summarizes the final composition of the cross-cancer dataset used for the experiments.

Table 5.16: Composition of the cross-cancer dataset after subsampling.

Cancer Type	Tumor Samples	Normal Samples	Total Samples
Leukemia	1051	106	1157
Breast Cancer	1051	106	1157
Total	**2102**	**212**	**2314**

This tumour type balanced configuration allows evaluation of the proposed GRN-based classification framework in a setting where regulatory signals must be learned across distinct tissue contexts.

This dataset differs from the leukemia-only cohort in both biological scope and graph content, since the union of samples from multiple tissues may alter the inferred regulatory landscape and the resulting patient-specific network variability.

5.8.2 Hyperparameter Optimization

Following the strategy used in the leukemia cohort, hyperparameter optimization was performed directly on the selected GCN2 architecture. The search focused on the most relevant training parameters, including hidden

dimension, dropout, optimizer settings, class weighting, and loss function. Table 5.13 summarizes the best-performing parameter set obtained during model selection, also for the cross-cancer dataset.

The selected configuration was then used for the final evaluation through cross-validation.

5.8.3 Final Cross-Validation Results

The final GCN2 model was evaluated using cross-validation on the cross-cancer GRN dataset. Table 5.17 reports the per-fold and average classification results.

Table 5.17: Final cross-validation results of the selected GCN2 model on the cross-cancer patient-specific GRN dataset.

Fold	Accuracy	Precision	Recall	Macro F1	Balanced Acc.
1	0.421	0.964	0.378	0.543	0.617
2	0.471	0.958	0.437	0.600	0.623
3	0.488	0.989	0.440	0.610	0.697
4	0.484	0.946	0.457	0.616	0.601
5	0.701	0.938	0.719	0.814	0.621
Average	**0.513**	**0.959**	**0.486**	**0.637**	**0.632**

5.8.4 Cross-Cancer Classification with HVG-Filtered Balanced Dataset

To further investigate the impact of feature selection and class balance in the cross-cancer disease association setting, an additional dataset variant was constructed using a High-Variable-Gene (HVG) filtering strategy combined with class rebalancing. The goal was to reduce transcriptomic noise while preserving the most informative expression signals for tumour detection.

Starting from the original cross-cancer cohort, the tumour class was downsampled using the same centrality-based selection strategy previously adopted during dataset construction. This resulted in a more balanced dataset composed of 212 normal samples and 1050 tumour samples.

In addition, a variance-driven feature selection step was applied by retaining the 2000 most highly expressed genes across the cohort, described in section 3.3.2. After alignment with transcription factor (TF) binding motifs and regulatory priors used in the PANDA/LIONESS pipeline, the resulting

dataset contained 2791 genes across 1262 samples. Table 5.18 summarises the main dataset properties.

Table 5.18: Statistics of the HVG-filtered cross-cancer dataset.

Property	Value
Total samples	1262
Normal samples	212
Tumour samples	1050
Selected highly variable genes	2000
Final genes after TF/motif alignment	2791

The regulatory priors were subsequently reconstructed using the same PANDA and LIONESS workflow applied throughout the thesis. Despite the reduced gene set, the resulting regulatory graphs remained structurally rich, containing millions of transcription factorâgene interactions. The statistics of the inferred networks are reported in Table 5.19.

Table 5.19: Regulatory network statistics for the HVG cross-cancer dataset.

Component	Value
Motif unique TFs	639
Motif unique genes	863
Motif interactions	4,803
PPI interactions	18,911
Final TF nodes (LIONESS graphs)	4,618
Final gene nodes	1,217
TF-gene interactions	5,620,106
Positive interactions	669,566

Graph neural network models were trained using the same experimental pipeline employed in the previous cross-cancer experiments. Notably, the HVG-based preprocessing resulted in improved classification performance compared to all previously tested dataset configurations.

Table **??** summarises the cross-validation results obtained using the best-performing configuration.

Table 5.20: Final cross-validation results of the selected GCN2 model on the HVG-balanced cross-cancer patient-specific GRN dataset.

Fold	Accuracy	Precision	Recall	Macro F1	Balanced Acc.
1	0.723	0.973	0.686	0.804	0.796
2	0.763	0.957	0.748	0.840	0.792
3	0.738	0.934	0.738	0.824	0.738
4	0.778	0.953	0.771	0.853	0.790
5	0.774	0.958	0.762	0.849	0.798
Average	**0.755**	**0.955**	**0.741**	**0.834**	**0.783**

Interestingly, the HVG filtering strategy produced the opposite effect compared to the leukemia-only experiments, where same process is investigated but no improvements were found. In the leukemia classification task, reducing the gene space using highly variable genes led to a slight degradation in predictive performance, likely due to the removal of lineage-specific markers that are critical for distinguishing AML and ALL subtypes.

In contrast, in the cross-cancer disease association setting the HVG preprocessing improved model performance. This behaviour can be explained by the different nature of the classification task. While leukemia subtype discrimination depends on subtle transcriptional differences between related hematopoietic lineages, tumour-versus-normal classification across tissues benefits from focusing on the most variable transcriptional programs associated with oncogenic transformation.

Finally, the lower feature dimensionality also produced more compact node embeddings. The resulting representation space will be visualised using UMAP projections to highlight the separation between tumour and normal samples. Fig. 5.11

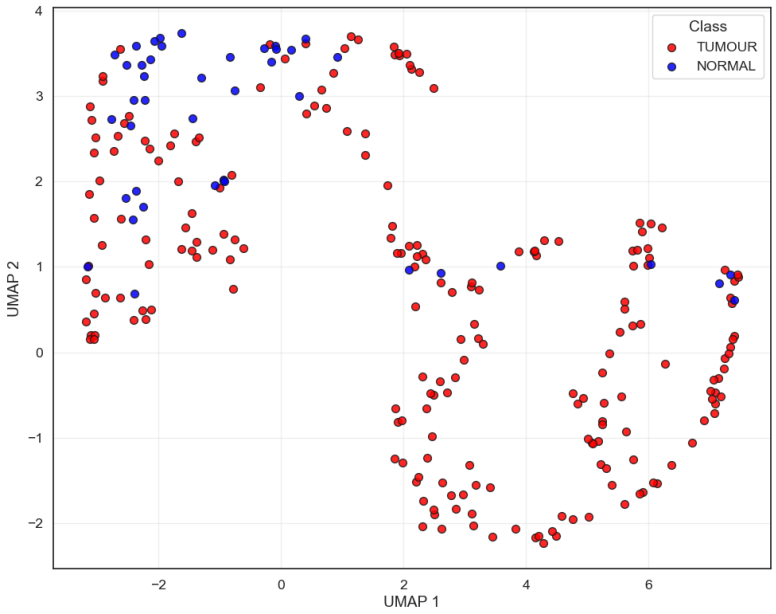


Figure 5.11: UMAP projection of graph embeddings learned from patient-specific GRNs in the cross-cancer setting. Each point represents one sample-specific GRN colored by class label.

5.8.5 Cross-Cancer Classification Discussion

The cross-cancer experiment resulted in an average balanced accuracy of 0.632 and a macro F1-score of 0.637. Compared with the leukemia-only classification task, the overall accuracy is lower (0.513), reflecting the increased difficulty of learning regulatory patterns that generalize across distinct cancer types and tissue contexts.

Interestingly, the model achieved very high precision (0.959) but considerably lower recall (0.486). This behavior indicates that when the model predicts a specific class it is usually correct, but it fails to identify a substantial portion of positive cases. Such a pattern suggests a conservative decision boundary, where the model relies on highly confident regulatory signals while ignoring weaker or more heterogeneous patterns across patients.

Several factors may explain the reduced performance relative to the leukemia-only setting. First, the cross-cancer regulatory network is substantially larger and denser. The GRN contains approximately 46 million edges compared to about 28 million in the leukemia-only network, and includes a larger number of genes (5836 vs 4797) and unique transcription factors (7877 vs 5837). While this richer network structure potentially captures more regulatory relationships, it also introduces additional noise and increases the complexity

of the learning task.

Second, the class distribution is considerably more imbalanced in the cross-cancer dataset. The tumor-to-normal ratio is approximately 1051:106 per cancer type, whereas the leukemia-only dataset has a milder imbalance (2136 tumor samples vs 175 normal samples). Such imbalance can further bias the classifier toward conservative predictions and contributes to the observed gap between precision and recall.

Despite these challenges, the balanced accuracy above 0.63 indicates that the GRN-based representation still captures regulatory signals that partially generalize across cancer types. These results suggest that patient-specific regulatory networks contain biologically meaningful information that can support cross-disease learning, although the heterogeneity of cancer regulatory programs and the increased network complexity remain limiting factors.

To further investigate the impact of feature selection and dataset balance, an additional cross-cancer dataset was constructed using a High-Variable-Gene (HVG) filtering strategy combined with a partial rebalancing of the tumour class.

Interestingly, this HVG-balanced configuration produced a clear improvement in predictive performance, reaching an accuracy of 0.755, a balanced accuracy of 0.783 and a macro F1-score of 0.834. Compared with the original cross-cancer dataset, the model shows both higher recall and improved balanced accuracy, while maintaining high precision. This suggests that reducing transcriptomic dimensionality and partially correcting class imbalance helps the GNN focus on stronger regulatory signals and reduces noise in the inferred regulatory graphs.

This behavior contrasts with the leukemia-only experiments, where HVG filtering slightly degraded performance. A plausible explanation is that leukemia subtype classification depends on subtle lineage-specific markers that may be removed by aggressive variance-based filtering, whereas tumour-versus-normal discrimination across tissues benefits from focusing on the most variable transcriptional programs associated with oncogenic transformation.

Table 5.21: Comparison between leukemia-only and cross-cancer classification using the final GCN2 model on patient-specific GRNs. Imbalance Ratio (IR) computed as N_{max}/N_{min}.

Setting	Acc.	Bal. Acc.	Macro F1	Imb. Ratio
Leukemia multiclass	0.611	0.696	0.588	4.71
Cross-cancer binary set 1	0.513	0.632	0.637	19.83
Cross-cancer HVG balanced	0.755	0.783	0.834	4.95

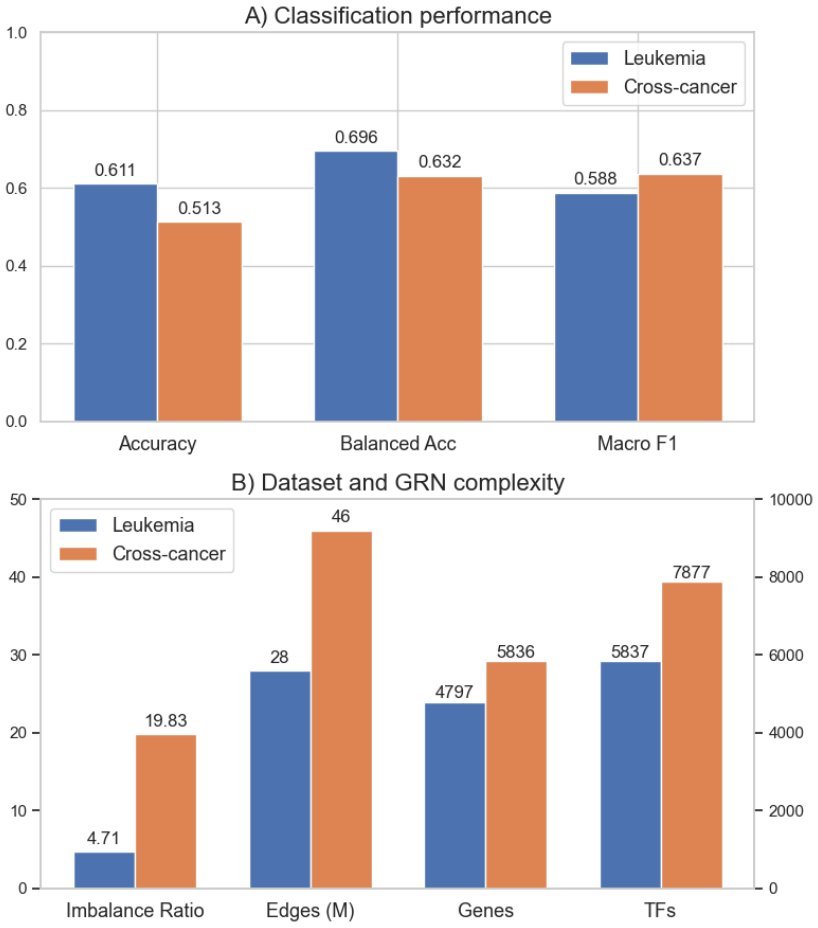


Figure 5.12: Comparison between the leukemia-only and cross-cancer experiments. (A) Classification performance of the final GCN2 model on patient-specific GRNs. (B) Structural and statistical characteristics of the datasets, highlighting the stronger class imbalance and the increased regulatory network complexity in the cross-cancer setting.

5.9 Explainability

To interpret the predictions of the GNN models, two explainability approaches were applied on GRN-based graphs: a SHAP-like feature attribution method and GNNExplainer. The analyses were conducted on both the leukemia dataset and the cross-cancer dataset.

Also the patient similarity graph models were analysed but following the procedure of integrated gradients.

5.9.1 Leukemia Dataset

SHAP-like Explanations on GRN Graphs

To investigate which regulatory components contributed most to the model predictions, a SHAP-like attribution analysis was performed on the patient-specific GRN graphs. The global feature importance was computed separately for each leukemia class and the top ranked genes were compared across classes.

Interestingly, the first 50 ranked genes were identical for the three classes (Normal, ALL, and AML), with differences observed only in the attribution magnitude rather than in gene identity. This indicates that the model relies on a largely shared set of regulatory signals across leukemia subtypes, rather than strongly class-specific biomarkers. The corresponding attribution heatmap is shown in Figure 5.13.

Despite the absence of class-specific rankings in the top features, several genes among the highest-ranked ones have known associations with leukemia biology. For instance, *ZRSR2* is a splicing factor frequently mutated in myeloid malignancies and acute myeloid leukemia, where alterations in RNA splicing programs contribute to leukemogenesis [60]. Similarly, members of the ATP-binding cassette transporter family such as *ABCG2* and *ABCC1* are well known mediators of multidrug resistance in acute leukemias and have been associated with treatment response and disease prognosis [76, 84].

The presence of these genes among the highest-ranked features suggests that the model captures biologically relevant regulatory programs associated with leukemia, although the lack of class-specific differentiation among the top-ranked genes reflects the substantial overlap between transcriptional programs of leukemia subtypes.

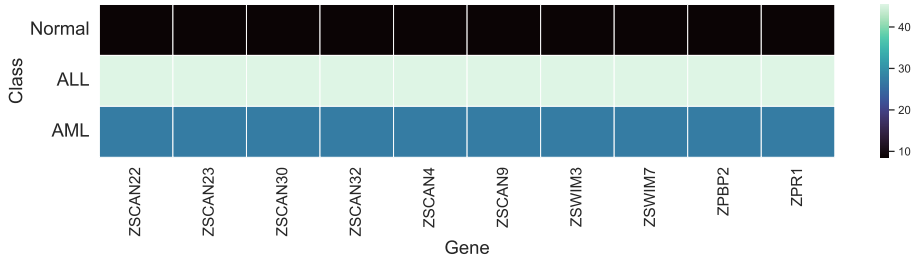


Figure 5.13: Heatmap of top gene importance scores identified by the SHAP-like explainer for each leukemia class.

Saliency Maps and Integrated Gradient Explanations on Patient Similarity Graphs

Top genes from patient similarity graphs. Explainability analysis on the patient–patient similarity graph was conducted using gradient-based attribution methods, specifically saliency maps and integrated gradients. For each patient, the most influential genes were first identified at the node feature level and subsequently aggregated across samples to obtain global class-specific importance rankings.

The overlap analysis of the top 50 genes per class reveals a strong class-specific signal. Only one gene appears among the highest ranked genes across all three classes, while the majority of genes are unique to individual leukemia subtypes. The ALL class contains several genes associated with lymphoid development, including *RAG1*, *DNTT*, *PTCRA*, and *VPREB1*, which are well-established markers of immature B-cell populations and early lymphoid differentiation. In contrast, the AML class is dominated by genes related to granulocyte and neutrophil activity such as *ELANE*, *MMP8*, *MMP9*, and *CEACAM8*, reflecting transcriptional programs associated with myeloid lineage differentiation.

To improve interpretability, the most discriminative genes across classes were further selected by measuring the attribution gap between the dominant class and the second highest attribution score, Fig 5.14. The resulting heatmap highlights a clear separation between lymphoid- and myeloid-associated transcriptional signatures, with ALL-specific markers clustered separately from AML-related genes. This pattern indicates that the patient similarity graph representation primarily captures lineage-specific transcriptional programs directly encoded in the gene expression profiles of the samples.

This behavior differs from the explainability results obtained from GRN-based models, where shared regulatory hubs tend to appear across leukemia

subtypes due to the network-level representation of transcriptional regulation. In contrast, the similarity graph operates directly on expression-derived features, making it more sensitive to lineage-defining gene expression signals rather than global regulatory topology. Such differences between feature-centric and network-centric representations are consistent with previous studies showing that expression-based models often highlight lineage markers, while network-based approaches capture broader regulatory programs underlying cancer heterogeneity [102, 105].

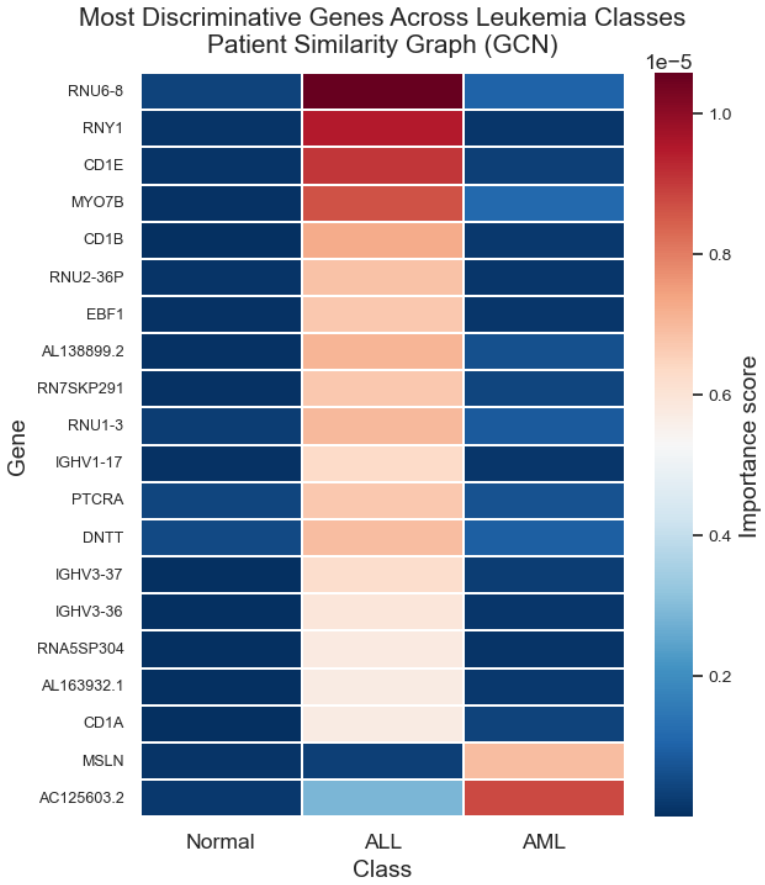


Figure 5.14: Heatmap of the most discriminative genes identified through saliency maps and integrated gradients in the patient similarity graph model. Genes are ordered according to the class in which they show the highest attribution score. The resulting pattern highlights lineage-specific transcriptional programs, with lymphoid markers enriched in ALL and myeloid markers enriched in AML.

5.9.2 Cross-Cancer Dataset

Explainability on the HVG Balanced Cross-Cancer Dataset

Explainability analysis for the cross-cancer task was conducted only on the HVG-balanced dataset. The original cross-cancer configuration achieved a relatively low balanced accuracy (0.63), which could lead to misleading interpretations of feature importance. In contrast, the HVG-balanced dataset produced substantially stronger predictive performance (balanced accuracy > 0.75), suggesting that the learned decision boundaries are more reliable and that the extracted explanations are more robust.

Furthermore, the SHAP-like explainability method implemented in this work computes feature attributions only for correctly classified samples. This strategy ensures that explanations are derived from true positive predictions rather than from potentially noisy or incorrect model outputs.

For the final analysis, same amount of graphs were explained both for normal and tumour samples. For each graph, gene-level contributions were computed and subsequently aggregated to obtain class-specific importance rankings based on the mean absolute attribution values.

Top genes for the normal class. The genes identified for the normal class are largely associated with hematopoietic and immune-related biological processes. Among the most influential genes are *MPO*, *S100A8*, *MMP8*, *ELANE*, and *CEACAM8*. These genes are known markers of neutrophil activity and granulocyte differentiation, processes that are frequently altered in hematological malignancies.

For instance, *MPO* encodes myeloperoxidase, a well-established marker of myeloid lineage cells and a key diagnostic marker in acute myeloid leukemia (AML) [2]. Similarly, *S100A8* and *S100A9* are inflammatory proteins involved in innate immune responses and have been reported as regulators of tumor-associated inflammation and leukemic microenvironments [31]. Proteases such as *MMP8* and *MMP2* participate in extracellular matrix remodeling and have been linked to cancer progression and tumor invasion [104]. The presence of these genes suggests that the model captures biological processes associated with immune surveillance and hematopoietic differentiation.

Top genes for the tumour class. In contrast, the tumour class is characterized by genes associated with epithelial and oncogenic transcriptional programs. Highly ranked genes include *KRT7*, *KRT19*, *GATA3*, *TFF1*, and *CRABP2*. These genes are frequently reported in epithelial cancers and are particularly associated with breast cancer transcriptional signatures.

Keratin genes such as *KRT7* and *KRT19* are structural proteins commonly used as epithelial tumor markers and have been widely reported in breast carcinoma studies [10] . *GATA3* is a key transcription factor regulating luminal epithelial differentiation and is one of the most established biomarkers in breast cancer pathology [54]. Similarly, *TFF1* (Trefoil Factor 1) is an estrogen-responsive gene strongly associated with luminal breast cancer subtypes and hormone receptor signaling pathways.

Other genes such as *ERBB2*, *ESR1*, and *MUC1* also appear among the highly ranked features, further supporting the interpretation that the model captures canonical oncogenic pathways associated with breast tumor biology.

Overall, the explainability analysis suggests that the model differentiates between normal and tumour samples by relying on biologically meaningful transcriptional programs. The normal class is characterized by genes related to hematopoietic and immune functions, while the tumour class is dominated by epithelial and breast cancer-associated transcriptional markers. This behavior is consistent with previous findings showing that transcriptomic classifiers tend to identify lineage-specific markers and oncogenic transcriptional programs when distinguishing between healthy and malignant tissues [96, 102].

Visualization of class-specific genes. To provide an intuitive representation of the most influential genes, Figure 5.15 presents a heatmap summarizing the top-ranked genes for both classes based on the aggregated SHAP-like attribution scores.

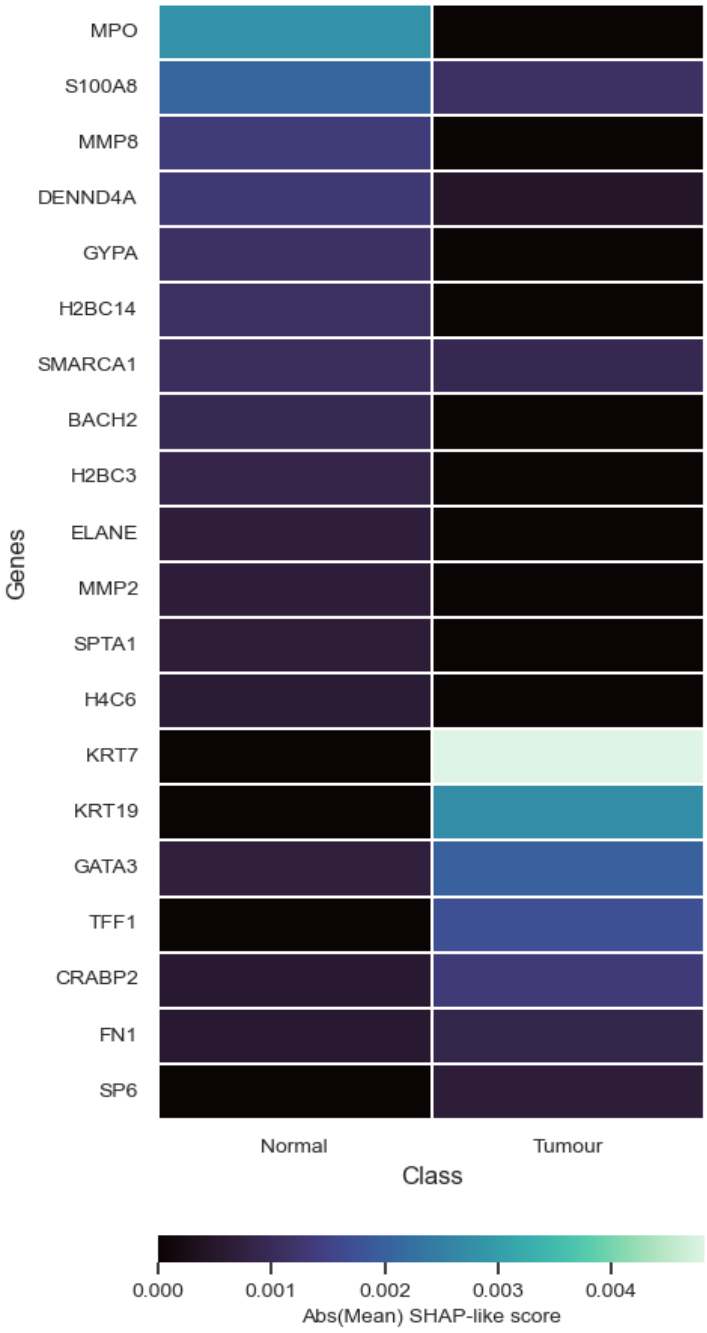


Figure 5.15: Heatmap of the most influential genes identified by the SHAP-like explainer on the HVG-balanced cross-cancer dataset. Colors represent the mean absolute attribution score aggregated across correctly classified samples for each class.

GNNExplainer Pathway-Level Results on the HVG-Balanced Cross-Cancer Dataset

Pathway-level explainability with GNNExplainer was performed only on the HVG-balanced cross-cancer dataset. As discussed in the previous section.

The pathway analysis was conducted separately for the two classes using Hallmark gene sets, from the Molecular Signatures Database (MSigDB) [53]. The Hallmark collection summarizes well-defined biological processes by integrating multiple gene expression signatures into coherent pathways, thereby reducing redundancy and improving interpretability. Many of these pathways are directly relevant to cancer biology, including processes such as epithelial–mesenchymal transition, hypoxia, inflammatory signaling, cell-cycle regulation, and p53-mediated stress responses. These mechanisms are widely implicated in both hematological malignancies such as leukemia and solid tumors including breast cancer, making the Hallmark gene set collection particularly suitable for cross-cancer transcriptomic analysis.

The resulting rankings show a remarkable degree of overlap between classes, with nearly identical top pathways and very similar ordering. In both classes, the highest scoring pathways include *TNFα signaling via NFκB*, *epithelial–mesenchymal transition*, *estrogen response early*, *G2M checkpoint*, *hypoxia*, *p53 pathway*, and *apoptosis*. This strong concordance suggests that GNNExplainer is primarily identifying global biological programs that are highly relevant to the tumour-versus-normal decision boundary, rather than sharply class-specific mechanisms. Such behaviour is plausible in a cross-cancer setting, where the model is expected to capture broad oncogenic and inflammatory processes shared across tumour states.

Among the most prominent pathways, *TNFα/NFκB signaling* is particularly relevant. NFκB activation is widely implicated in cancer-associated inflammation, tumour cell survival, and resistance to therapy, and has also been specifically linked to both breast cancer progression and acute myeloid leukemia biology [28, 90]. Similarly, *epithelial–mesenchymal transition* is a hallmark of tumour plasticity, invasiveness, and metastatic potential, especially in breast cancer, where EMT-related transcriptional programs are strongly associated with aggressive phenotypes [39]. The prominence of *estrogen response* pathways is also biologically coherent, given the central role of estrogen receptor signaling in breast cancer subtype definition and tumour progression [37].

Additional highly ranked pathways, including *hypoxia*, *p53 pathway*, *apoptosis*, and *G2M checkpoint*, are consistent with canonical tumour-associated programs involving metabolic stress, defective cell-cycle control, genomic instability, and altered survival signaling [2, 99]. At the same time, immune-

related pathways such as *IL2/STAT5 signaling*, *interferon gamma response*, *complement*, and *coagulation* indicate that the model may also be exploiting inflammatory and microenvironment-related signals, which are often shared across malignant tissues and systemic hematologic alterations [62].

Overall, these results support the view that the GNN classifier relies on biologically meaningful pathway-level signals. However, because the pathway rankings are highly similar between the two classes, the findings should be interpreted mainly as evidence of robust global cancer-related programs rather than as strictly class-specific pathway biomarkers. In this sense, the GNNExplainer results complement the SHAP-like gene-level analysis: while SHAP-like explanations highlighted more interpretable gene-level differences between normal and tumour samples, GNNExplainer emphasized higher-level pathways that appear to organize the global structure of the decision function.

Figure 5.16 reports the top pathways identified in the two classes, for example as a compact heatmap or as two aligned horizontal bar plots.

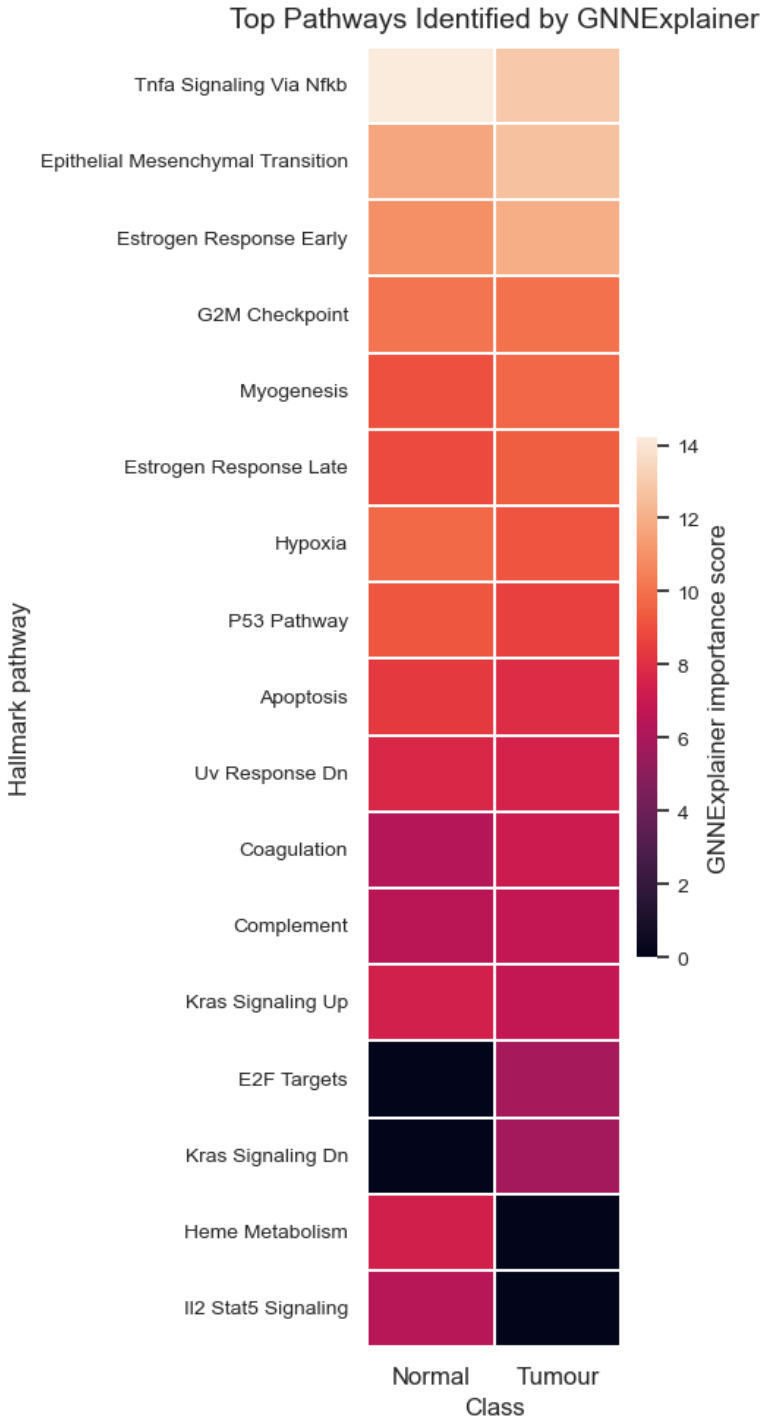


Figure 5.16: Top pathways identified by GNNExplainer on the HVG-balanced cross-cancer dataset. Scores represent the importance of each pathway in the model decision process.

Chapter 6

Conclusions and Future Directions

6.1 Overview of the Study

The rapid expansion of high-throughput transcriptomic technologies has transformed computational oncology by enabling the systematic analysis of gene expression profiles across large patient cohorts. RNA sequencing (RNAseq) captures the functional state of cells and provides a quantitative view of transcriptional activity under diverse physiological and pathological conditions [16,46]. Large repositories such as the Genomic Data Commons (GDC) and The Cancer Genome Atlas (TCGA) have further accelerated the development of ML methods for cancer classification and biomarker discovery.

Despite these advances, most transcriptomic classification pipelines treat genes as independent predictors. This assumption conflicts with established principles of systems biology and network medicine, according to which biological phenotypes arise from coordinated interactions among molecular components rather than isolated gene alterations [11]. In cancer, dysregulation often manifests as a rewiring of regulatory and signalling networks, affecting transcriptional programs that govern cell proliferation, differentiation, immune response, and tissue-specific phenotypic states.

Motivated by this systems-level perspective, this thesis investigated whether integrating **gene regulatory networks** with **Graph Neural Networks** can improve the modelling of inter-patient heterogeneity in leukemia and, more broadly, support cross-cancer disease-association analysis. To address this question, a structured computational pipeline was developed combining transcriptomic preprocessing, regulatory network inference, graph-based ML, and explainable artificial intelligence techniques.

Regulatory networks were inferred using the PANDA algorithm, which integrates gene expression with transcription factor binding priors and protein–protein interaction information to estimate regulatory interactions [33]. To capture patient-level variability, the LIONESS framework was applied to reconstruct single-sample regulatory networks, enabling the extraction of individual regulatory landscapes from cohort-level models [48].

Two complementary modelling paradigms were explored. First, **gene expression-based models (GEX models)** operate directly on transcriptomic matrices using dimensionality reduction and classical ML algorithms. Second, **graph-based models** exploit structural relationships among biological entities by representing transcriptomic data as networks and applying GNNs for classification.

Beyond predictive modelling, particular emphasis was placed on **explainability**, which is essential for the adoption of machine learning in biomedical contexts. Explainable AI techniques enable the identification of biologically relevant features driving model predictions, helping bridge the gap between predictive accuracy and mechanistic understanding [102].

Finally, to evaluate the generalisability of the proposed framework, the modelling pipeline was extended to a *cross-cancer disease-association setting*, integrating leukemia and breast cancer transcriptomic datasets. In the final stage of the work, this setting was further refined through a **highly variable gene (HVG)-filtered and class-balanced dataset**, which substantially improved the reliability of the graph-based cross-cancer analysis and enabled a more robust interpretation of both gene-level and pathway-level explanations.

6.2 Summary of Main Findings

The results obtained in this study highlight the strengths and limitations of different modelling paradigms for transcriptomic classification.

Gene expression-based models achieved the highest predictive performance overall. After dimensionality reduction using principal component analysis (PCA), high-dimensional RNAseq matrices were transformed into compact feature representations suitable for ML algorithms. Classical models and shallow neural networks demonstrated strong classification performance in both binary and multiclass leukemia classification tasks. These findings confirm previous studies showing that gene expression profiles alone provide strong discriminative signals for leukemia subtype classification [19, 79].

Graph-based approaches using **patient–patient similarity networks** also produced competitive predictive performance. In this setting, patients

are represented as nodes connected according to similarity in their transcriptomic profiles. GNNs applied to these networks were able to capture relational patterns among samples and achieve high classification accuracy. Explainability techniques such as saliency maps and integrated gradients further identified sets of genes contributing to patient-level predictions, revealing subtype-specific transcriptional signatures.

In contrast, **regulatory network-based models constructed from PANDA and LIONESS** showed lower predictive performance. The extremely large size and density of regulatory graphs, often containing millions of transcription factor–gene interactions, introduce substantial computational and optimisation challenges for GNN training.

However, regulatory network modelling provided a key advantage: **biological interpretability**. Analysis of PANDA-derived regulatory networks revealed transcription factors and regulatory hubs whose activity differed across disease conditions. Several transcription factors exhibited differential regulatory influence between tumour and normal samples, consistent with the network medicine view that cancer arises from perturbations in molecular interaction networks [11].

A particularly relevant result of this thesis is that the behaviour of GRN-based models depended strongly on the dataset formulation and on the nature of the classification task. In the leukemia-only setting, HVG filtering did not improve predictive performance, likely because subtype discrimination between AML and ALL depends on subtle lineage-specific transcriptional programs and on markers that may be lost through aggressive feature reduction. By contrast, in the **cross-cancer tumour-versus-normal setting**, the HVG-filtered and class-balanced dataset led to a clear improvement in graph-based classification performance. The best GNN configuration on the HVG-balanced cross-cancer dataset achieved an average balanced accuracy of approximately **0.78**, substantially improving over the previous cross-cancer setting and showing that task-oriented feature selection and class rebalancing can make patient-specific GRN learning more effective.

This result is important because it indicates that GRN-based graph models are not intrinsically weak, but rather highly sensitive to data curation, graph size, class imbalance, and the biological structure of the task. Tumour-versus-normal classification across tissues appears to benefit from focusing on the most variable transcriptional programs associated with oncogenic transformation, whereas leukemia subtype classification requires preserving more fine-grained lineage-related information.

Explainability analyses further reinforced these findings. For leukemia, the most interpretable regulatory insights were obtained directly from the PANDA-derived networks rather than from the GRN-based GNN classifiers,

whose explanations remained comparatively limited. For the final cross-cancer HVG-balanced experiment, instead, both **SHAP-like** and **GNNExplainer** analyses produced biologically coherent signals.

At the gene level, SHAP-like explanations highlighted a biologically meaningful contrast between the two classes. Normal samples were mainly associated with hematopoietic and immune- related markers such as *MPO*, *S100A8*, *MMP8*, *ELANE*, and *CEACAM8*, whereas tumour samples were driven by epithelial and breast-cancer-related genes such as *KRT7*, *KRT19*, *GATA3*, *TFF1*, *CRABP2*, *ERBB2*, *ESR1*, and *MUC1*. These patterns are consistent with the biological contrast between hematopoietic normal tissues and tumour samples enriched for epithelial programs [2, 10, 31, 54, 96, 102].

At the pathway level, GNNExplainer identified global Hallmark pathways that were highly relevant to the tumour-versus-normal decision boundary, including *TNFα signaling via NFκB*, *epithelial–mesenchymal transition*, *estrogen response early*, *G2M checkpoint*, *hypoxia*, *p53 pathway*, and *apoptosis*. These pathways are strongly implicated in cancer-associated inflammation, tumour plasticity, hormone-driven oncogenic programs, defective cell-cycle control, and stress adaptation [28, 37, 39, 53, 62, 90, 99]. Although the pathway rankings were highly similar across classes and therefore should be interpreted as global oncogenic programs rather than strictly class-specific biomarkers, they provide evidence that the model captures coherent higher-level biological processes.

Overall, the cross-cancer disease-association experiments demonstrated that the modelling pipeline can generalise beyond leukemia-specific tasks. More importantly, the HVG-balanced variant showed that when the dataset is better aligned with the classification objective, patient-specific GRN-based GNNs can achieve substantially improved performance together with more reliable and interpretable explanations.

6.3 Methodological Contributions

This thesis provides several methodological contributions at the intersection of computational biology, graph ML, and explainable AI.

First, a scalable pipeline for **patient-specific regulatory graph modelling** was developed. By integrating PANDA-based regulatory inference with LIONESS single-sample reconstruction, it was possible to generate large collections of patient-specific regulatory networks suitable for downstream ML analyses.

Second, the work presents a **systematic comparison of modelling paradigms** for transcriptomic classification, including:

- feature-based models operating on gene expression matrices,
- GNNs applied to patient similarity graphs,
- GNNs trained on regulatory graphs derived from GRN inference.

This comparative analysis highlights the trade-offs between predictive accuracy, interpretability, and computational complexity when modelling large-scale biological networks.

Third, the study integrates multiple **explainability techniques** to interpret model predictions in biological terms. Feature attribution methods including SHAP-like approaches, saliency maps, and integrated gradients were applied to gene expression and similarity-based models, while GNNExplainer was explored for graph-based classifiers. In the final cross-cancer HVG-balanced setting, these explainability approaches became substantially more informative, enabling both class-specific gene interpretation and pathway-level biological analysis. Explainable AI methods are increasingly recognised as essential tools for understanding ML models in biomedical research [102].

Finally, the modelling framework was extended to a **cross-cancer disease-association paradigm**, demonstrating the potential of network- based approaches to capture shared oncogenic signals across tissues. The introduction of an HVG-filtered and balanced graph dataset further showed that **dataset engineering is not merely a preprocessing detail, but a central methodological component** for improving both predictive performance and explanation quality in GRN-based GNN frameworks.

6.4 Limitations of the Study

Several limitations of the present study should be acknowledged.

First, the transcriptomic datasets used in this work exhibit **class imbalance**, with certain leukemia subtypes represented by substantially larger numbers of samples than others. Although balancing strategies were applied during model training, residual imbalance may still influence predictive performance.

Second, regulatory network inference introduces methodological assumptions that may affect downstream analyses. PANDA integrates prior information from transcription factor binding motifs and protein–protein interaction networks, which themselves may contain incomplete or noisy information.

Third, the **size and density of LIONESS-derived networks** represent a major computational challenge. Single-patient regulatory graphs may

contain millions of edges, making them difficult to process efficiently using current graph neural network architectures.

Fourth, explainability methods for GNNs remain an active research area. Although feature attribution methods provided meaningful insights for gene expression and for the final cross-cancer HVG-balanced experiment, the interpretability of graph-based explanations remains limited, particularly when attempting to map local graph explanations to canonical biological pathways in a class-specific manner. In the cross-cancer setting, GNNExplainer mostly highlighted broad, shared tumour-related pathways rather than sharply class-discriminative mechanisms.

Finally, the cross-cancer disease-association analysis included only one additional tumour type. Further validation across larger multi-cancer cohorts will be necessary to fully assess the robustness of the proposed framework and the transferability of the learned regulatory programs.

6.5 Future Research Directions

The results of this study open several promising directions for future research.

First, improved **data curation and integration** could further enhance modelling performance. Future work could incorporate larger transcriptomic cohorts and apply advanced batch-correction techniques to mitigate technical variability. The improvement observed in the HVG-balanced cross-cancer setting suggests that more systematic studies on balancing strategies and task-specific feature filtering would be particularly valuable.

Second, exploring **alternative regulatory network inference methods** may lead to more suitable graph representations. Algorithms such as GENIE3, GRNBoost, or LogBTF may produce sparser or more context-specific regulatory networks that could be more amenable to graph-based learning [52].

Third, research into **graph simplification strategies** may help address the computational challenges associated with large regulatory graphs. Techniques such as network sparsification, edge filtering, or biologically informed pruning could reduce graph complexity while preserving key regulatory interactions.

Fourth, advances in **GNN architectures** may improve the modelling of large biological networks. Emerging models such as graph transformers may offer improved scalability and expressiveness, particularly for heterogeneous and multi-resolution biological graphs [104].

Finally, further work is needed to improve **explainability methods for graph-based models**. Developing approaches that link graph explanations

to biological pathways, regulatory modules, and tissue-specific gene programs would significantly enhance the interpretability of GNNs in biomedical applications. In particular, combining gene-level attribution and pathway-level graph explanations in a unified framework may provide a more complete view of how graph models organize biological decision boundaries.

6.6 Final Remarks

In summary, this thesis investigated the integration of gene regulatory networks and GNNs for transcriptomic cancer modelling. While gene expression-based models remain the strongest solution for predictive classification in the leukemia setting, regulatory network representations provide complementary insights into the biological mechanisms underlying disease.

The results also show that the usefulness of GRN-based graph learning depends strongly on how the dataset is defined. In the leukemia-only setting, GRN-based GNNs remained mainly valuable as a framework for regulatory interpretation through PANDA network analysis. In the cross-cancer setting, however, the HVG-balanced dataset led to substantially better balanced accuracy and enabled more robust explainability, showing that patient-specific regulatory graph models can become more effective when graph size, class balance, and feature space are better aligned with the classification objective.

Although GNNs applied to large regulatory graphs remain computationally challenging, the combination of regulatory network inference and explainable ML represents a promising direction for computational oncology.

These findings reinforce the importance of combining predictive modelling with biologically meaningful representations. By bridging ML with systems biology, network-based approaches may contribute to more interpretable and robust models for understanding cancer heterogeneity, disease associations, and regulatory mechanisms.

Table 6.1: Comparison of modelling paradigms explored in this thesis.

Modelling paradigm	**Predictive performance**	**Biological interpretability**	**Computational complexity**
Gene expression models	High	Limited	Low
Patient similarity GNN	High	Moderate	Moderate
GRN-based models (PANDA/LIONESS) in leukemia	Moderate	High	High
GRN-based models (PANDA/LIONESS) in cross-cancer HVG-balanced setting	Moderate to high	High	High

Bibliography

[1]

[2] Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine*, 368(22):2059â2074, May 2013.

[3] The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 2020.

[4] The biogrid database: curated protein and genetic interactions. *Nucleic Acids Research*, 2021.

[5] Kegg: integrating viruses and cellular organisms. *Nucleic Acids Research*, 2023.

[6] String database in 2023: proteinâprotein association networks. *Nucleic Acids Research*, 2023.

[7] AIEOP. Incidence of acute lymphoblastic leukemia in children in italy, 2025.

[8] AIRC/AIRTUM. Leukemia incidence in italy by sex, 2025. Adult leukemia incidence estimates.

[9] M. et al. Almamun. Integrated methylome and transcriptome analysis reveals novel regulatory elements in pediatric acute lymphoblastic leukemia. *Epigenetics*, 2015.

[10] Konstantina Athanasopoulou, Vasiliki-Ioanna Michalopoulou, Andreas Scorilas, and Panagiotis G. Adamopoulos. Integrating artificial intelligence in next-generation sequencing: Advances, challenges, and future directions. *Current Issues in Molecular Biology*, 47(6):470, June 2025.

[11] Albert-Laszlo Barabasi, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, December 2010.

[12] Leo Breiman. Random forests. *Machine Learning*, 45(1):5â32, October 2001.

[13] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022.

[14] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36:411–420, 2018.

[15] R.R. et al. Canevarolo. The expression and activation of the nf-κb pathway in pediatric acute lymphoblastic leukemia. *Genes*, 2023.

[16] Geng Chen, Baitang Ning, and Tieliu Shi. Single-cell rna-seq technologies and related computational data analysis. *Frontiers in Genetics*, 10, April 2019.

[17] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *International Conference on Machine Learning*, 2020.

[18] J.H. et al. Cheng. $CamkiiÎ^3 regulatestheviabilityandself - renewalofacutemyeloidleukemiastem - likecells. CellDeathDiscovery$, 2021.

[19] X. et al. Cheng. Automated prediction of acute promyelocytic leukemia from flow cytometry data using a graph neural network pipeline. *Bioinformatics*, 2025.

[20] C. Cobaleda, A. Schebesta, A. Delogu, and M. Busslinger. Pax5: the guardian of b cell identity and function. *Nature Immunology*, 8:463–470, 2007.

[21] International Childhood Cancer Registry Consortium. Childhood leukemia: epidemiology and incidence, 2023. 25â30% of childhood cancers; 75% ALL.

[22] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273â297, September 1995.

[23] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[24] C.V. Dang. Myc on the path to cancer. *Cell*, 149:22–35, 2012.

[25] Benjamin DeMeo and Bonnie Berger. Hopper: a mathematically optimal algorithm for sketching biological data. *Bioinformatics*, 36(Supplement_1):i236–i241, 2020.

[26] C. et al. Desterke. Egr1 dysregulation defines an inflammatory and leukemic program in hematopoietic stem cells. *JCI Insight*, 2021.

[27] B. et al. Di Francesco. Nf-κb: A druggable target in acute myeloid leukemia. *Cancers*, 2022.

[28] Bruno Di Francesco, Giuseppina Todisco, Francesco Barile, et al. Nf-κb : $Adruggabletargetinacutemyeloidleukemia.\ Cancers, 14(15)$: 3557, 2022.

[29] TSGroup Politecnico di Torino. Gnn-diseases: Graph neural networks for disease module identification. https://gitlabtsgroup.polito.it/root/gnn-diseases, 2026. Accessed: 2026-03-15.

[30] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *AAAI*, 2021.

[31] Amgad Mohamed Elshoeibi, Ahmed Badr, Basel Elsayed, Omar Metwally, Raghad Elshoeibi, Mohamed Ragab Elhadary, Ahmed Elshoeibi, Mohamed Amro Attya, Fatima Khadadah, Awni Alshurafa, Ahmad Alhuraiji, and Mohamed Yassin. Integrating ai and ml in myelodysplastic syndrome diagnosis: State-of-the-art and future prospects. *Cancers*, 16(1):65, December 2023.

[32] B. et al. Fiordi. Il-18 and vegf-a trigger type 2 innate lymphoid cell accumulation and pro-tumoral function in chronic myeloid leukemia. *Leukemia*, 2023.

[33] Kimberly Glass, Curtis Huttenhower, John Quackenbush, and Guo-Cheng Yuan. Passing messages between biological networks to refine predicted interactions. *PLoS ONE*, 8(5):e64832, May 2013.

[34] G. Gogoshin and A. Rodin. Graph neural networks in cancer and oncology research: Emerging and future trends. *Cancers*, 2023.

[35] T. R. Golub, D. K. Slonim, and P. Tamayo. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[36] Robert L. Grossman, Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109â1112, September 2016.

[37] Jennifer M. Grunda, Amber D. Steg, Qing He, et al. Differential expression of breast cancer-associated genes between stage- and age-matched tumor and normal breast tissues. *BMC Research Notes*, 5:481, 2012.

[38] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, 2017.

[39] M. Haque et al. Targeted therapy approaches for epithelial-mesenchymal transition in triple-negative breast cancer. *Current Molecular Pharmacology*, 2024.

[40] Brian Hie, Hyunghoon Cho, Benjamin DeMeo, Bryan Bryson, and Bonnie Berger. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. *Cell Systems*, 8(6):483–493.e7, 2019.

[41] Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.e6, 2018.

[42] Weihua et al. Hu. Strategies for pre-training graph neural networks. *ICLR*, 2020.

[43] J. Huang et al. Disease burden, risk factors, and trends of leukaemia. *Journal of Clinical Medicine*, 2022. 474,519 new cases globally; 5.4 per 100,000 incidence.

[44] Y. A. et al. Huang. Graph structure learning for tumor microenvironment with cell type annotation from non-spatial scrna-seq data. *Bioinformatics*, 2024.

[45] E.A.R. et al. Ismail. Birc6/apollon gene expression in childhood acute leukemia. *American Journal of Hematology*, 2012.

[46] D. et al. Jovic. Single-cell rna sequencing technologies and applications. *Nature Communications*, 2021.

[47] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.

[48] Marieke L. Kuijjer, Minggang Tung, Guang Yuan, John Quackenbush, and Kimberly Glass. Estimating sample-specific regulatory networks using lioness. *iScience*, 14:226–240, 2019.

[49] R. et al. Kulkarni. Early growth response factor 1 in aging hematopoietic stem cells and leukemia. *Frontiers in Cell and Developmental Biology*, 2022.

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

[51] S. et al. Laukkanen. Six6 is a tal1-regulated transcription factor associated with t-all. *Blood Advances*, 2020.

[52] L. Li, Z.P. Liu, et al. Logbtf: Logistic regression estimation-based boolean threshold function for grn inference. *Bioinformatics*, 39(5):btad256, 2023.

[53] Arthur Liberzon, C. Birger, H. ThorvaldsdÃ^3*ttir, et al.Themolecularsignaturesdatabaseha* 417 − −425, 2015.

[54] Henrik Lilljebjörn, Christina Orsmark-Pietras, Felix Mitelman, Anna Hagström-Andersson, and Thoas Fioretos. Transcriptomics paving the way for improved diagnostics and precision medicine of acute leukemia. *Seminars in Cancer Biology*, 84:40â49, September 2022.

[55] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, 2017.

[56] Orsolya Liska, Balázs Bohár, András Hidas, Tamás Korcsmáros, Balázs Papp, Dávid Fazekas, and Eszter Ari. TFLink: an integrated gateway to access transcription factorâtarget gene interactions for multiple species. *Database*, 2022:baac083, 2022.

[57] Meng Liu, Hongyang Gao, and Shuiwang Ji. Towards deeper graph neural networks with elastic graph neural networks. *arXiv preprint arXiv:2107.07999*, 2021.

[58] Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019.

[59] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *NeurIPS*, 2017.

[60] V. Madan and colleagues. Zrsr2 mutations in myeloid malignancies. *Leukemia*, 2020.

[61] D. et al. Moujalled. Targeting apoptosis in acute myeloid leukemia. *British Journal of Cancer*, 2017.

[62] Noor S. Naji et al. Inflammation and related signaling pathways in acute myeloid leukemia. *International Journal of Molecular Sciences*, 2024.

[63] National Cancer Institute. Citing target in publications and presentations. Web page, 2022. Accessed: YYYY-MM-DD.

[64] National Cancer Institute. Therapeutically applicable research to generate effective treatments (target). Web page, 2022. Accessed: YYYY-MM-DD.

[65] National Cancer Institute. Genomic Data Commons Data Portal. https://portal.gdc.cancer.gov/, 2024. Accesso effettuato il: 14 marzo 2026.

[66] National Cancer Institute Genomic Data Commons. Gdc data portal: TCGA-BRCA project (breast invasive carcinoma). https://portal.gdc.cancer.gov/projects/TCGA-BRCA. Accessed 2026-02-28.

[67] T. Pabst, B.U. Mueller, and P. et al. Zhang. Dominant-negative mutations of cebpa in acute myeloid leukemia. *Nature Genetics*, 27:263–270, 2001.

[68] Siavash Pai and Gary Bader. netdx: patient classification using integrated patient similarity networks. *Nature Methods*, 2014.

[69] Roberto Passera, Giulia Zamagni, Elisa Fabbro, Giulia Carreras, Caterina Ledda, Carlo La Vecchia, Lorenzo Giovanni Mantovani, Andrea Maugeri, Mario Virgilio Papa, Dimitri Poddighe, Silvano Gallus, Pawan Sirwan Faris, Mihajlo Jakovljevic, Giuseppe Minervini, Giuseppe Gorini, and Lorenzo Monasta. The national burden of leukemia in italy from 1990 to 2023: results from the global burden of disease study 2023. *eClinicalMedicine*, 88:103509, October 2025.

[70] A. et al. Paszke. Pytorch documentation. https://pytorch.org/docs, 2023.

[71] J.A. et al. Pulikkan. Cell-cycle regulator e2f1 and microrna-223 comprise an autoregulatory negative feedback loop in acute myeloid leukemia. *Blood*, 2010.

[72] D. et al. Raman. Role of chemokines in tumor growth. *Cancer Letters*, 2007.

[73] Ricardo Ramirez, Yu-Chiao Chiu, et al. Classification of cancer types using graph convolutional neural networks. *Frontiers in Physics*, 2020.

[74] R.G. Ramsay and T.J. Gonda. Myb function in normal and cancer cells. *Nature Reviews Cancer*, 8:523–534, 2008.

[75] P. et al. RimmelÃ©. The spi1/pu.1 transcription factor accelerates replication fork progression and promotes preleukemic cell proliferation. *Cell Reports*, 2017.

[76] R. et al. Robey. Abcg2: a perspective. *Advanced Drug Delivery Reviews*, 2018.

[77] Enakshi Saha, Viola Fanfani, Panagiotis Mandros, Marouen Ben-Guebila, Jonas Fischer, Katherine Hoff-Shutta, Kimberly Glass, Dawn L. DeMeo, Camila M. Lopes-Ramos, and John Quackenbush. Bayesian inference of sample-specific coexpression networks. *Genome Research*, 34(7):1114–1125, 2024.

[78] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33:495–502, 2015.

[79] Sharanya Selvaraj, Alhuseen Omar Alsayed, Nor Azman Ismail, Balasubramanian Prabhu Kavin, Edeh Michael Onyema, Gan Hong Seng, and Arinze Queen Uchechi. Super learner model for classifying leukemia through gene expression monitoring. *Discover Oncology*, 15(1), September 2024.

[80] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

[81] Riccardo Smeriglio, Joana Rosell-Mirmi, Petia Radeva, and Jordi Abante. Leveraging protein-protein interactions in phenotype prediction through graph neural networks. August 2024.

[82] Dongyuan Song, Nan Mu Xi, Jessie J. Li, and Lulu Wang. scsampler: fast diversity-preserving subsampling of large-scale single-cell transcriptomic data. *Bioinformatics*, 38(11):3126–3127, 2022.

[83] R. Sood, Y. Kamikubo, and P. Liu. Role of runx1 in hematological malignancies. *Blood*, 129:2070–2082, 2017.

[84] D. et al. Steinbach. Abcc1 (mrp1) expression and multidrug resistance in leukemia. *Blood*, 2002.

[85] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019.

[86] Ming Tang, Å½eljko AntiÄ, Pedram Fardzadeh, Stefan Pietzsch, Charlotte Schröder, Adrian Eberhardt, Alena van Bömmel, Gabriele Escherich, Winfried Hofmann, Martin A. Horstmann, Thomas Illig, J. Matt McCrary, Jana Lentes, Markus Metzler, Wolfgang Nejdl, Brigitte Schlegelberger, Martin Schrappe, Martin Zimmermann, Karolina Miarka-Walczyk, Agata Pastorczak, Gunnar Cario, Bernhard Y. Renard, Martin Stanulla, and Anke Katharina Bergmann. An artificial intelligence-assisted clinical framework to facilitate diagnostics and translational discovery in hematologic neoplasia. *eBioMedicine*, 104:105171, June 2024.

[87] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[88] J. et al. UngerbÃ¤ck. The role of tcf3 in b-cell acute lymphoblastic leukemia. *Leukemia*, 29:1843–1853, 2015.

[89] Petar et al. Velickovic. Graph attention networks. *ICLR*, 2018.

[90] Wei Wang, Sanchita A. Nag, and Run Zhang. Targeting the nfκ*bsignalingpathwaysforbreastcancerpreventionandtherapy.CurrentMedicinalChemis* $264 - -289, 2015$.

[91] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. M. Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[92] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *ICLR*, 2019.

[93] Chaoyi Yin, Yangkun Cao, et al. Molecular subtyping of cancer based on robust graph neural network and multi-omics data integration. *Frontiers in Genetics*, 2022.

[94] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *NeurIPS*, 2019.

[95] Tianwei Yue, Yuanxin Wang, Longxiang Zhang, Chunming Gu, Haoru Xue, Wenping Wang, Qi Lyu, and Yujie Dun. Deep learning for genomics: A concise overview, 2018.

[96] Tianwei Yue, Yuanxin Wang, Longxiang Zhang, Chunming Gu, Haoru Xue, Wenping Wang, Qi Lyu, and Yujie Dun. Deep learning for genomics: From early neural nets to modern large language models. *International Journal of Molecular Sciences*, 24(21):15858, November 2023.

[97] Q. Zhang et al. Hallmarkgraph: a cancer hallmark informed graph neural network for hierarchical tumor subtype classification. *Bioinformatics*, 2025.

[98] Y. et al. Zhang. Sp1 and c-myc modulate drug resistance of leukemia stem cells by regulating survivin expression through the erk-msk mapk signaling pathway. *Molecular Cancer*, 2015.

[99] Shuqin Zhi et al. Hypoxia-inducible factor in breast cancer: role and target for therapy. *Journal of Translational Medicine*, 2024.

[100] C. et al. Zhou. Jun is a key transcriptional regulator of the unfolded protein response in acute myeloid leukemia. *Leukemia*, 2017.

[101] Jie Zhou et al. Graph neural networks: A review of methods and applications. *AI Open*, 2020.

[102] Zhongliang Zhou, Mengxuan Hu, Mariah Salcedo, Nathan Gravel, Wayland Yeung, Aarya Venkat, Dongliang Guo, Jielu Zhang, Natarajan Kannan, and Sheng Li. Xai meets biology: A comprehensive review of explainable ai in bioinformatics applications, 2023.

[103] Marinka Zitnik et al. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 2018.

[104] Mohammad Zohari et al. Graph neural networks in multi-omics cancer research. *Briefings in Bioinformatics*, 2024.

[105] Payam Zohari and Mostafa Haghir Chehreghani. Graph neural networks in multi-omics cancer research: A structured survey, 2025.