Generalized Bridge Pipeline for Multimodal Gene Regulatory Network Discovery

by

Abrar Sami Khan Alif
24141153
Riyadus Salehin Fahmid
21201205
Mohammad Omar Raihan
21141058
Omor Bin Amjad Chowdhury
23241085

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
October 2025.

© 2025. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Abrar Sami Khan	Riyadus Salehin Fahmid
24141153	21201205
Mohammad Omar Raihan	Omor Bin Amjad Chowdhury
21141058	23241085

Approval

The thesis/project titled "Generalized Bridge Pipeline for Multimodal Gene Regulatory Network Discovery" submitted by

1. Abrar Sami Khan Alif (24141153)
2. Riyadus Salehin Fahmid (21201205)
3. Mohammad Omar Raihan (21141058)
4. Omor Bin Amjad Chowdhury (23241085)

of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in October 9, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

Dr. Swakkhar Shatabda

Professor
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Dr.Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor & Chairperson
Department of Computer Science and Engineering
BRAC University

Abstract

In this thesis, a computationally generalized pipeline where multimodal gene regulatory networks (GRNs) are built by combining transcriptomic data in RNA sequencing and functional dependency data in CRISPR knockout screens. Traditional forms of GRN rely on expression data as the only tool which can not detect causal or functional significant interactions. We attempted to solve this by constructing a contrastive bridge model, where both datasets are put in the same 128-dimensional latent space. We used Maximum Mean Discrepancy (MMD) loss and a diversity-preserving loss such that patterns of modality are aligned, and meaningful biological variation is not distorted. Using these embeddings, we built multimodal GRNs, combining evidence as provided by various outlets. In order to identify statistical and functional relationships, we demonstrated Spearman co-expression correlations, GENIE3 random forest importance scores, CRISPR dependency support, and cosine similarity of embedding vectors into a single edge-weight expression. The bridge-fused networks have been steady in structure and introduced new cross-modal interactions (Bridge-fused vs GENIE3) when used in both hematopoietic and lung cell data. Top hub genes in these networks scored negative on the mean CRISPR dependency score which is an indication of important functional roles and Gene Ontology enrichment analysis scored significant representation of the immune activation and metabolic processes. The implications of these findings are that the bridge pipeline offers biologically meaningful, consistent and interpretable GRNs. Overall, this framework is a generalizable and data-driven framework to integrate heterogeneous genomic datasets, which can be applied in the process of identifying significant regulators and potential therapeutic targets in a broad variety of biological settings.

Keywords: Gene Regulatory Network, Multimodal Integration, Contrastive Learning, Maximum Mean Discrepancy, Co-expression, CRISPR Dependency, Functional Genomics, Network Inference, Cross-Modal Alignment, Gene Prioritization.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

GRNs represent the interactions between gene's with the genes interactions and regulation being the nodes and edges respectively[3], [33]. These networks constitute the fundamental conceptualization in computational systems biology since they help in explaining how cells differentiate, respond to external stimuli, and develop disease. GRN inference GRN inferences Traditional GRN inference is the use of single-modality expression data (microarray or RNA-seq) to infer potential relationships[5]. The early approaches constructed co-expression graphs with an edge between two genes denoting that the two genes were extremely correlated in their sample expression. Correlation is useful but fails to give a clear cut on supporting direct regulation and indirect effects due to a shared pathway or confounding factors.

The simple correlation was later improved with additional algorithms such as the ARACNe, CLR and GENIE3, incorporating information and machine learning techniques[5], [7]. In the GENIE3, special care is taken to use groups of regression trees to estimate how the expression of a gene contributes to predicting another, given the feature-importance scores as a measure of regulation. These best-performing methods, in the DREAM challenges, showed that, even in the case of only expression data, aspects of the underlying regulatory structure were partially known[7]. However, expression-only approaches are usually noisy and incomplete because many of the co-expressed genes are not necessarily related, as not all of the true-regulatory relationships are necessarily manifested by strong expression correlations in all situations[11].

To overcome this disadvantage, there are more recent studies that are based on multimodal data integration[24], [37]. The reasoning is that the different forms of experimental data possess varying evidence: in case two genes are truly related, their association must be represented more than once. CRISPR/Cas9 perturbation screens, such as those generated by DepMap, are a great complementary source of data in genomes[9], [16]. This is experimentally determined by knocking out each gene and assessing the viability of hundreds of human cell lines, resulting in a dependency profile of each gene. Analysis of such profiles gives co-essentiality: genes with a similar viability profile in all cell lines are likely to be operative in the same functional module or pathway[9]. Although the directionality of regulation is not

mentioned in CRISPR screens, they reveal functional connectivity, which cannot be characterized by the expression[16].

Such a mixture of RNA-seq and CRISPR data will combine observational and interventional data[27], [30]. True relationships can be demonstrated either in the form of correlated expression and shared knock-out phenotypes[36]. The fact that these evidences coincide contributes even further to the assumption that the relationship is biologically meaningful[33]. To forecast a consolidated and more confident GRN, our work goes one step further by proposing a single multimodal architecture known as the Bridge pipeline, which trains a shared representation of genes across modalities[24], [37]. This is followed by the check of the resulting network by the Gene Ontology (GO) enrichment analysis[2], [4], [6], [21] which makes sure that the optimal hubs in the network are not a haphazard graph structure but a steady biological system.

1.2 Rational of the Study or Motivation

In case of gene regulation, single-modality analyses don't provide us with the full story[3]. CRISPR-based co-essentiality do a great job of showing functional groupings but fall short on explaining the regulatory mechanisms of direction, relying solely on expression-based networks can lead to misleading correlations[16]. Potential causal relationships can be hinted by expression data alongside providing context, while solid evidence of functional connections can be offered by dependency data: the pipeline takes advantage of the best of both worlds by combining these two approaches[9].

From a computational point of view, aligning representations from two different views of the same entities is the goal of this challenge[24], [37]. Each gene exists in two feature spaces—one that captures its expression behavior across various cell populations and another that reflects its functional essentiality in cancer cell lines. Despite being having two different modalities, out objective was to align them so that they can be compared in a common format. An agreement between transcriptional activity and functional dependency can be revealed through simple similarity calculations in the shared space, making it easier to identify biologically relevant modules and hubs[36].

To make it possible, the bridge network uses two encoders that map expression-based and CRISPR-based features into a common 128-dimentsional latent space. To align the overall distributions, the training process used Maximum Mean Discrepancy (MMD), while using diversity-preserving regularizer to maintain gene variance and prevent collapse. Neither contrastive positive-negative pair loss nor PCA was needed as the encoder itself learns to compress high-dimensional features into the 128-D space. Besides offering flexibility, this design allows the same framework to work with additional omics modalities in future research.

1.3 Problem Statement

Despite having had an abundance of genomic data, it is still tough to piece together Gene Regulatory Networks (GRNs)[3], [33]. This is mainly because each data comes with its own set of biases and thus gives us only a glimpse of the whole picture[11]. We examine whether combining gene-expression data with CRISPR screen gene-dependency data is able to improve regulatory discovery in this subsection[16]. This challenge is decomposed into several sub-tasks: how to make a statistically sound alignment between gene-level of the two types of data and which kind of evidence will be put together to form a combined network?[24] that neither type would dominate the other in terms contributions[36] with the final question being if indeed, an improvement in reliability was achieved; for instance comparing SNP pair accuracy from networks.

On the derivation of the RNA space and corresponding artificial neural network model that generates similar low-dimensional embeddings for each gene both in data types, this translates into the task of training a neural network to generate similar such embeddings in data type A and B.[24] We then compute single and multimodal edge scores, which quantify the probability of a functional or regulatory link, respectively[5], [7]. It then evaluates whether these scores retain identifiable structure and biologically relevant hubs[33]. We validate our improvements using a variety of quantitative measures for the performance, such as the Jaccard index for measuring network overlap, Kendall's τ score for measureing the stability of hub rankings[1] and the mean CRISPR support_z parameter (`meanZ`) which gauges functional validation[16], along with biological coherence measured via Gene Ontology (GO) enrichment of top hubs[2], [4].

The ultimate goal is to develop a Bridge + fusion pipeline, which inputs two data types for the same genes and generates a high-quality GRN[24], [37]. This approach is also designed to mitigate any distributional mismatches while capturing and fusing informative signals from the two distinct datasets[36].

1.4 Objective

1.4.1 Developing the Bridge Alignment Model

The main aim was to design a model that could learn shared latent space for gene features from both CRISPR and RNA-seq methods. Two encoder branches for two different modalities was used to design the bridge network, generating 128-D embeddings. During training, our aim was to bring the composite loss that includes the MMD term as down as possible, as this helps align the distributions of the two sets of embeddings. Gene variance was maintained by diversity-preserving regularizer that prevents the embeddings from collapsing into a single point. Reducing dimensionality was internally handled by the model. Ultimately, the resulting latent space captured the relationships between different modalities while still maintaining the unique characteristics of each gene.

1.4.2 Constructing the Multimodal Fusion Framework

The second part of the pipeline was to build a Gene Regulatory Network using various sources of evidence. For each pair of genes, the system calculated four complementary metrics: the absolute Spearman correlation of their expression profiles, the GENIE3 random-forest feature-importance score that reflects regulatory influence, a CRISPR-based co-essentiality measure adjusted to a standardized support z, and the cosine similarity of the genes from the bridge embeddings. A weight fusion formula was used to combined these metrics, where support mean represents the average of the normalized CRISPR support z and the normalized cosine similarity. Additionally, to reduce the weight of edges that show disagreement across different modalities, cosine similarity was used as a refinement step, ensuring that the final network highlights consistent evidence.

CRISPR Fusion:

$$\text{Edge Weight} = \alpha|\rho| + \beta \cdot R_{\text{imp}} + (1 - \alpha - \beta) \cdot \text{support}_{mean}$$

Bridge Fusion:

$$\text{Edge Weight} = \alpha|\rho| + \beta \cdot R_{\text{imp}} + (1 - \alpha - \beta) \cdot cosine\_similarity$$

1.4.3 Evaluating Network Performance

The third step was to evaluate the performance and effectiveness of the integrated network compared to single-modality baselines. The analysis focuses on quantitative stability and functional significance. The assessment focuses on quantitative stability and functional significance. We assess structural similarity by calculating Jaccard index of the common edges and analyze the preservation of important regulators using Kendall's τ correlation between hub rankings. To validate the biological legitimacy, on which we calculate the mean support_z of our top hubs (negatively numbers indicate to essential genes) and perform GO enrichment analyse for these hubs to guarantee that they are co-contributed in immune and metabolic pathways. All in all, these metrics let us decide whether the Bridge pipeline increases reliability without creating noise.

1.4.4 Building a Generalizable and Reproducible Pipeline

One of our aims is to develop a flexible and reusable computational workflow that takes care of the whole chain from preprocessing data, training models, fusing evidences and evaluating models. These steps are articulated in a well-documented code that can be repeated speaking about other datasets or tissues only with some minor modifications. We've explicitly specified hyper-parameters such as fusion weights (α, β) and embedding dimension, ensuring our work would be able to be reproduced and extended by others in studying other modalities.

1.5 Methodology in Brief

There are a series of steps withing the overall framework. We collected CRISPR gene-dependency data from DepMap. We also gathered gene-expression data focus-

ing on hematopoietic populations from Human Cell Atlas, and pseudobulk expression profiles from Lung Atlas. The datasets had different gene naming conventions, so we had to harmonize all gene identifiers using GHNC-approved mappings to ensure everything aligned perfectly accross the two different modalities. The CRISPR dependency scores were standardized so that any negative values indicated lethality when a gene is knocked out. The expression values were also log-normalized and scaled. For the sake of the analysis, genes expressing extremely low variance and information were left out.

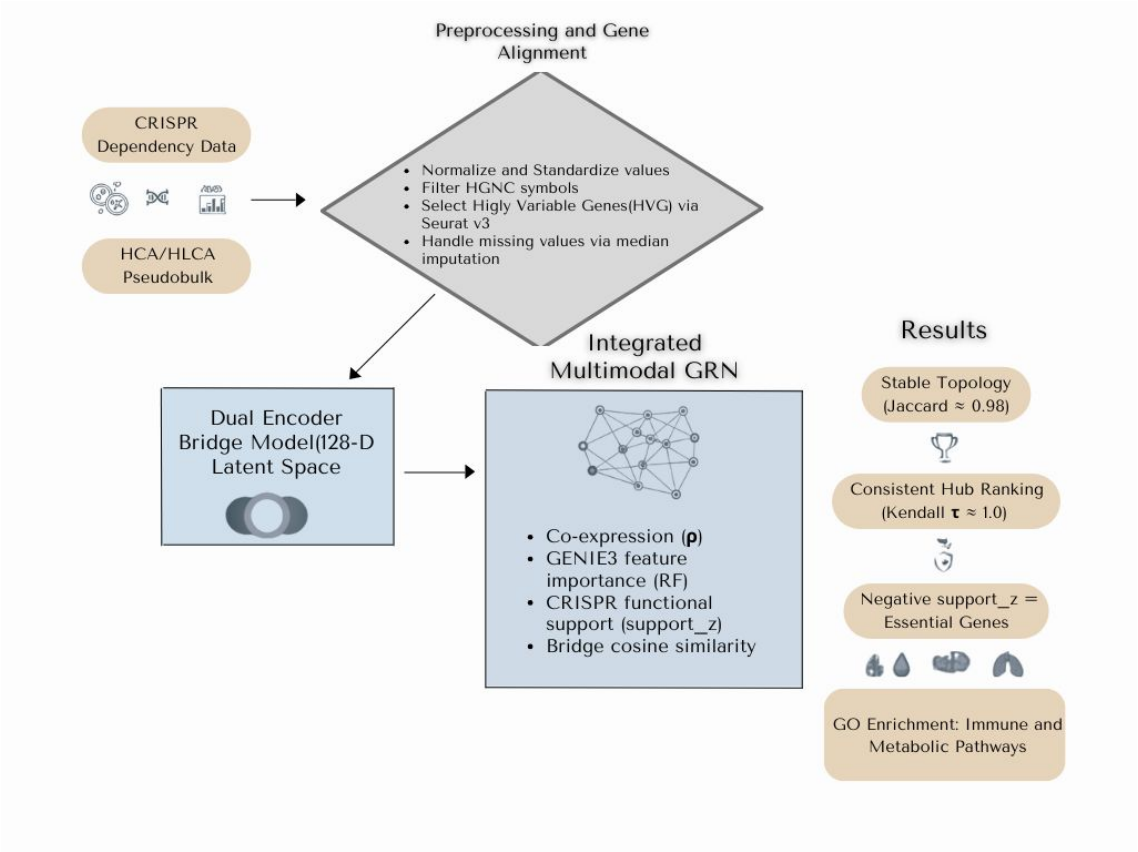

Figure 1.1: Implementation Workflow

Each gene was represented by two input vectors: one that captured its expression profile and another that reflected its dependency profile. These vectors were fed into the Bridge network's two encoders. The encoders performed dimensionality reduction internally, learning a 128-dimensional latent representation through MMD and diversity losses.

Once the bridge was done with training, it generated a unique embedding for each gene. We calculated various evidence scores for all gene pairs by leveraging the combination of the embeddings and the original data. The scores included the absolute Spearman correlation derived from expression data, GENIE3 feature-importance values, CRISPR co-essentiality adjusted to normalized support z, and the cosine similarity between gene embeddings. All of these were integrated using the weighted formula mentioned earlier to produce a single multimodal edge score. Cosine similarity acted both as a contributing feature and as a filter to help eliminate inconsistent edges. The gene pairs were then ranked based on their fused scores, and a prede-

termined number of the highest-scoring pairs made up the final undirected network. Finally, we evaluated how the bridge integrated networks differed from the baseline networks.

1.6 Scopes and Challenges

This project mainly focused on developing, implementing, and validating a flexible computational pipeline for discovering multimodal gene regulatory networks (GRNs). We used two biological contexts-hematopoietic and lung systems to showcase our approach, but it's designed to work with any combination of modalities that share a common gene set. Rather than seeking complete biological discovery, we prioritized internal consistency and credibility and demonstrated success through quantitative and enrichment-based validation.

We encountered a number of challenges during the development. The heterogeneity of data demanded careful normalization due to the difference in scale and dynamics of expression and CRISPR datasets, for which we used MMD loss to align their distributions in an embedding space. We have effectively reduced the problem of processing high-dimensional input data to 128 dimensions in the encoder, thus avoiding an external PCA. Alignment was unsupervised, and this fact threatened to cause collapse, which we countered with a fixed variance-preserving regularizer. To address the problem of threshold selection, we utilized statistical filters (e.g. FDR on co-expression) along with fixed edge count that would lead to similar values in order to make unbiased comparisons. We controlled for computational costs by training in mini-batches and tuning parameters carefully. Lastly, interpreting latent features can be tricky, and so we used hub-gene essentiality and enrichment analysis to validate the learned representations, helping deducing biological coherence without needing to directly interpret the latent dimensions.

In summary, the pipeline is a generalized and repeatable data-driven framework for integrating expression and perturbation data. It successfully established stable, biologically supported gene-regulatory networks and provided a flexible template for multimodal integration in future systems in biological studies.

Chapter 2

Literature Review

2.1 Preliminaries

Gene regulatory networks (GRNs) are often abstracted as directed graphs that visualize the interactions of regulators (genes or more generally their products such as transcription factors (TFs)) and their target genes [13]. In this network, each node corresponded to a gene, and a directed edge from one gene A to another B indicated that the former can regulate the latter (e.g., changing A's activity can modulate B's expression) [13]. For instance, transcriptional regulatory networks describe how transcription factors, enhancers along with other regulatory elements shape gene expression patterns in a cell-context specific manner [13], [14]. Combining different data types can help to capture this regulation: gene expression (e.g. RNA-seq, microarrays), chromatin state information (e.g. ChIP-seq or ATAC-seq) and results of perturbation experiments (such as knockdowns or CRISPR) all contribute valuable data about the connections in the network [13], [14]. In other words, we can have a more complete and informative GRN with the combination of multiple sources of evidence [14].

- **Graph representation:** Gene Regulatory Network (GRN) is a graph which can be defined as $G = (V, E)$, where each node V represents a gene; while the edges E represent the interactions between genes. An edge $(i \rightarrow j)$, in some sense, Mean that gene i is connected in a way to gene j and exercise by influence gene$i$ expects gene$j$. These edges in bayesian networks can also contain a probability or weight, which expresses the level of evidence for that regulatory relationship [13], [35]. Such networks can indeed mirror regulatory patterns occurring in some biological setting as they are tissue or condition specific [14].

- **Data sources:** Gene Regulatory Networks (GRNs) are the networks where each node represents a gene and is connected by edges if there is some kind of regulatory relation between them [14]. In order to obtain their expressions we can exploit statistics such as Pearsons correlation or regression fit, which then provides an insight on potential regulatory links [14]. Many analyses pipelines also incorporate prior knowledge such as transcription factor binding sites from motif databases or ChIP-seq experiments, protein–protein interaction networks, pathway annotations, or even chromatin accessibility data [14]. Large-scale experimental data that carry strong causal signal, such as genetic

perturbation through knockdown or CRISPR experiments (if knocking out gene A consistently changes the expression level of B (through some kind of measure), it is a strong indicator of a directed edge A→B in the network [13], [14]), can also have an enormous impact on function inference. Despite the fact that, in reality, to boost real interactions various evidence streams are used and combined together where the noise needs to be filtered in an effective GRN inference [13], [14].

- **Inference models:** There are various computational techniques that are used to infer Gene Regulatory Networks (GRNs). Information-theoretic approaches (such as mutual information), dependency network models, Bayesian networks (often acyclic), and Gaussian graphical models are some of the statistical techniques[14]. For instance, a typical two-step process involves first identifying potential regulator-target pairs, often with the help of known motifs or databases and then refining or scoring these pairs based on expression data[14]. During this phase, evaluation metrics like Pearson's correlation or regression coefficients are incorporated to assess the strength of each proposed connection**Baur**. However, regulators that lack known motifs may get overlooked if solely relied on motif filtering. Therefore, many methods incorporate additional strategies beyond just sequence priors[14].

- **Machine-learning approaches:** Nowadays, machine learning is often relied upon for to grasp complex non-linear or combinatorial regulation in Gene Regulatory Netowrks (GRNs). Using tree-based ensemble method like Random Forest is a standout approach as it plays a vital role in GENIE3 algorithm. In this method, each gene's expression is treated as a response variable, while the expressions of all other genes serve as inputs. The Random Forest additionally computes feature-importance scores, which assess the importance of each gene (the inputs) as a predictor from the target gene expression[5]. This will, for each possible regulator-target pair, produce a score or weight. Ensemble methods are attractive in that they can deal with high-dimensional data with many genes, model nonlinear interactions and generate easy-to-interpret importance rankings lacking in large parameter tuning[5]. More machine learning models for network inference such as Support Vector Machines, regression models or deep learning— have additionally been studied in parallel using both supervised and unsupervised approaches[14], [35].

- **Integrative frameworks:** Incorporating different kinds of data into a joint inference procedure is the main point in latter studies. For example, the iRafNet model incorporates a variety of prior data such as protein-protein interactions, TF binding assays and knockdown data in addition to gene expression data and IETFs while improving on the Random forest method. In practice, as opposed to relying on expression data alone, iRafNet uses these prior information to bias trees building or to modify feature weights, leading to more accurate TF–target predictions[10]. There are also Bayesian and supervised-learning methods that couple information from multiple sources. One such approach is an supervised model that, given gene expression and binding profiles, ontologies and known pathways compute the priors of edges[13]. Indeed, these instances share the concept of reinforcement of the true signal by giving more

weight to edges that are supported by many independent data sources[10], [13]. Our pipeline implements this idea by combining correlated-based, feature importance score based and external support measures to infer edge weights in a manner analogous to these existing integrative approaches.

- **Perturbation-based inference:** Finally, by some carefully designed experiments, we are able to recover GRN edges. As a reviewer notes: "When we do something to gene A in an experiment (e.g., knock down or out, or overexpression of it), and that manipulation results in some change in the expression of gene B, there is good reason to believe that the former regulates the latter[13]. This is not just an observed correlation: because we know what caused the change (the change in A), we can determine the direction of the relationship. There are several algorithms tailored to process such perturbation data to identify the directed edges. For example, CRISPR knockout screens have been recently used to construct networks (i.e., such as knocking out various transcription factors and observing changes to the level of transcripts in order to create trans-regulatory cascades connecting regulators with genes found through GWAS studies)[32]. In conclusion, perturbation assays are an efficient approach to reveal the structure of GRNs in combination with observation data[13], [32].

The existing literature demonstrates that for sound network inference one should be integrating among graph-theoretic models, statistical measures and machine-learning techniques while considering distinct kinds of data. These basic tenets become the mainspring of our work and in combination with co-expression signals, weight rankings as well as perturbance support our endeavor is directed towards building up a weighted gene network[5], [7]. In the remaining sections we will consider more closely our particular pipeline and how it extends these basic ideas.

2.2 Review of Existing Research

A range of methods starting from simple statistical techniques to sophisticated machine learning techniques exist to infer Gene Regulatory Networks (GRNs)[5], [7]. We have reviewed the main categories of these methods and the insights that inspired the design of the Bridge model.

- **Coexpression and information-theoretic methods:** At the onset of Gene Regulatory Network (GRN) inference, researchers primarily depended on observational gene expression data. Their go-to simple methods to construct co-expression networks were Pearson correlation and mutual information. For example, to find out potential regulatory connections, the ARACNe algorithm utilizes mutual information along with null hypothesis[7]. Meanwhile, WGCNA organizes genes into modules based on their correlation patterns[3]. These broad trends revealing are relatively easy to implement in gene coregulation computation. However, a significant downside is that correlationbased networks can mix up direct and indirect interactions. For instance, a misleading strong correlation might be caused by other intermediate regulators or a common upstream factor rather than directly related. This limitation means

that, using expression data alone will automatically generate networks with relatively high false positive signals.

- **Machine learning and ensemble approaches:** To escape from naïve heuristics, one possible way is to learn the gene regulatory network (GRN) from microarray data by machine. They have proved successful for regression-based techniques, like LASSO or elastic net, which infers a gene's expression using others', and tree-based ones, like GENIE3, applying random forests to rank regulators[5]. Each of these methods have their own aspects: GENIE3 is able to learn non-linear relationship by learning a set of decision trees, whereas LASSO supposes a linear relationship between the variables but produce easily interpretable sparse networks. Crucially, it has been demonstrated by the DREAM5 network inference challenge that there is no one-size-fits-all best method for all datasets[7]. Instead, a "wisdom of crowds" approach — averaging or otherwise combining predictions from many algorithms — gave the best predictions. They discovered that a consensus of 35 different algorithms performed better than individual ones, and the pooled model confirmed almost half of its predicted regulatory connections in E. coli experiments[7]. These community-based predictions during challenge 2 also indicated that the use of perturbation data, imposing expected network features (e.g., sparsity or modular organization), or incorporating prior knowledge can greatly improve the inference performance[11]. The important take-home message is: combining diverse sources of evidence (or inference methods) can greatly boost confidence in predicted networks. This observation directly influences our pipeline that combines several evidence metrics (correlation, feature importance, and CRISPR support) instead of relying on a single one.

- **Gene essentiality and functional genomics networks:** Now, depending on the rise of large-scale CRIPSR screening [9], [16], researchers have the capacity to construct drug networks around genetic dependency profiles. A good example of this are the co-essentiality networks. whereby genes with similar knockout fitness profiles in multiple cell lines become interconnected. Relationships between pathways or protein complexes are uncovered through these networks: when a group of genes shows strong dependencies in the same subset of cell lines, it indicates that they likely work together in a biological process. According to Jackson, co-essentiality-based gene clustering can help point out metabolic and signaling modules, shedding light on pathway structures that complement expression-based perspectives[26]. These CRISPR-derived networks are particularly effective for uncovering functional gene groupings[16]—like components of the same enzyme complex or parallel members of a pathway and they offer a more causal interpretation than coexpression networks. In this context, an edge signifies that impairing gene i and gene j leads to correlated effects on cell viability, implying that they share a common essential function. However, relying solely on co-essentiality networks might overlook regulatory interactions that don't show strong viability effects, and they can be swayed by global fitness factors (for example, genes that are universally essential or non-essential across cells will naturally correlate with many others). In out research, instead of using CRISPR co-essentiality in isolation, we integrated expression data with it. To propose that

authentic regulatory relationships will often represent both co-expression and co-essentiality signals, we aimed to apprehend the overlap of "co-regulation" and "co-functionality" by viewing the essentiality profile of each gene as another layer of data.

- **Multimodal data integration and latent-space models:** In the domain of computational biology, leveraging an integration of different types of data is slowly becoming influential[24], [37]. With an aim to capture signals common across all data types, various methods emphasizing on creating a shared low-dimensional representation have been established to integrate multi-omics data. Techniques like canonical correlation analysis (CCA) have been used to identify correlated latent variables between two datasets in the past. However, achieving complex non-linear alignments are much easier now with the rise of more flexible neural network models[21], [28], [34]. The recent advancements in domain adaptation and multi-view representation learning acts as an inspiration for our MMD-based shared latent space[24]. In domain adaptation like transferring knowledge from one species or experimental batch to another, it's common to align feature distributions from a "source" domain with a "target" domain to help a model generalize better. MMD has proven effective in aligning feature distributions within deep networks, ensuring that the learned embeddings remain consistent across different domains or data types[8], [24]. For example, to reduce distribution shifts between domains, Deep Adaptation Networks[8] incorporated MMD into a neural network, allowing for the learning of transferable features. Similarly, in generative modeling, Generative Moment Matching Networks (GMMNs) employed MMD as a training criterion to align generated samples with real data distributions. These examples support our method of using MMD to align gene expression and CRISPR essentiality data within a common latent space: we consider each data type as a "domain" and aim to find an embedding that makes the distributions of gene features as similar as possible. Unlike contrastive learning, which requires explicit positive/negative pair design or siamese networks, our MMD approach is purely distributional. It takes advantage of the fact that each gene has two perspectives (expression and dependency) by matching their overall feature statistics. This scheme has a solid existing foundation, thus simplifying the integration challenge.

- **Regularization to preserve variance and information:** One of the big challenges with unsupervised embedding methods, especially those that leverage powerful neural networks, is steering clear of degenerate solutions[33]. A model that's supposed to align two different modalities might end up collapsing into trivial embeddings, mapping all genes to the same point in latent space just to minimize a distance metric. Through distinct regularization of the embedding space, some solutions have been provided with the help of recent advancements in self-supervised learning. For instance, a redundancy-reduction loss that encourages the cross-correlation matrix of outputs from twin networks to resemble the identity matrix has been introduced by Barlow Twins[18]. This approach ensures that each latent feature carries unique information (driving off-diagonal correlations to zero) while keeping the two representations of each input closely aligned (with diagonal correlations approaching

one). Similarly, VICReg (Variance-Invariance-Covariance Regularization[20] introduced a term that guarantees each latent dimension has non-zero variance (which prevents it from collapsing to a constant) and a decorrelation term that minimizes covariance between different dimensions. Both of these diversity-preserving practices maintains a rich diversity in the data variance of the learned features and actively stops representations from collapsing. We want genes to take up different spots in the latent space so that they can demonstrate their distinctive regulatory roles. Thus, it is important for the gene embeddings to have diversity in terms of GRN inference. Our diversityp-reserving regularizer works in a similar way: it penalizes the model if all gene embeddings start to look too alike or if any latent dimension becomes trivial across genes. The quality and interpretatbility of the shared representation is upheld is enhanced by ensuring a rich and diverse structure in the latent space. The differences that set genes apart from one another are crucial for effectively distinguishing regulator and target genes. The Bridge model, aligns expression and CRISPR data without losing those subtle differences.

Prior research also set the stage for our approach in several important respects. Simple coexpression networks provide a good point of departure, but really worth every single hint support which may be available to them. Single-source methods are clearly dominated by ensemble and hybrid methods. Furthermore, we could avoid the drawback of mere correlational research by employing perturbation data such as CRISPR screening. Moreover, machine learning enables us to design strong multimodal embedding a tying techniques through inclusion of MMD for distribution matching while keeping (losing) diversity by regulazations. The Bridge model essentially unifies these ideas; that is, it connects two harmonizing data sources by accurate statistical alignment (MMD), and keep the integrated representation information-rich through variance-preserving regulation. As we would discuss, this design was trying to capture more reliable repetitive gene relationships based on biology that could not be completely reproduced by any single data type or a particular method.

2.3 Summary of Key Findings

From the studies we have done, there are a couple of key ideas that shape the design of our integrative GRN inference framework:

Need for Multi-Source Integration: The complexity inherent in GRNs is likely to be hidden from single data source or method. Although coexpression patterns can be inferred from expression data, they may not imply causality. By contrast, the CRISPR perturbation data might be more focused on viabilities instead of causal relationships. Integrated analysis of heterogeneous data adds depth to our understanding of these networks, while considering multiple types of evidence including expression data, predictive modeling and genetic perturbation is essential for achieving an accurate interpretation. This is reminiscent of the "wisdom of crowds" [7] in which pooling various methods result in more trustworthy networks.

- **Ensemble and Hybrid Methods Excel:** Different inference techniques can be exploited to increase performance. Our approach combined correlation

scores, machine learning-based feature importances, and co-essentiality signals to a unified edge scoring scheme[5], [11]. By giving the right importance to these features, we have limited their individual drawbacks and exploited their combined power in an evolutionary direction consistent wth the optimal solutions observed in previous benchmarks of phenomena inference for GRNs.

- **Value of Perturbation Data:** CRISPR co-essentiality networks combined with the predicted interactions significantly contributed to the functional validation of these associations. Genes with similar loss-of-function profiles are often involved in the same complexes or pathways 15,57,58. Incorporating the support from essentiality in our model, by grounding the network in experimental genetics evidence, and ascribing function to its edges increases confidence that these links reflect actual biology. This method enables us to identify, for instance, modules of mutually essential genes with potential regulatory complexes candidates[9], [16].

- **MMD for Cross-Modal Alignment:** The Maximum Mean Discrepancy (MMD) technique has received some popularity due to its success in matching distributions, particularly for machine learning and domain adaptation problems. By making one-to-one matching only based on gene identity instead of pairwise label, MMD is flexible and capable to non-parametrically learn the inferential mapping from two different kinds of data to a common latent space; appealingly, this alternative perfectly fits in distribution alignment. This allows the model to build a joint representation of genes, which accounts for both expression and dependency information. Therefore, it successfully addresses the complexity of the data and enables insights from both data types to contribute to network inference.

- **Avoiding Collapse with Regularization:** It is important that the learned structure of the latent space being informative and not cluttered with unimportant solutions. Lessons from self-supervised learning techniques such as Barlow Twins and VICReg suggest that we can obtain richer representations by introducing inductive biases to maintain variance under control and reduce redundancy in embeddings. Our diversity-preserving regularizer is based on these ideas, which in turn demands that the latent space retains the complexity of gene expression patterns and essentiality profiles. Such a step ensures the model retains significant discrepancies between gene profiles and does not "aver" out these important regulatory signals, so it is an essential step for downstream network construction.

- **Consistency with Biological Knowledge:** Real gene regulatory relationships frequently manifest through multiple evidence lines. For example, if a transcription factor is truly regulating its target gene, we would expect to observe them coexpressing under some conditions, high feature importance score linking them in a predictive model and similar essentiality profiles if it indeed is a critical interaction for cell survival. Once evidence is provided on all of these different dimensions, it highly indicates a true link in the network. The multimodal architecture of the Bridge model is designed to specifically capture this type of convergent evidence, increasing the chances that the inferred connections represent true biological interactions. By stipulating that

the expression-driven and dependency-driven signals agree to some extent, the model can effectively suppress spurious associations that appear in a single dataset while focusing attention on connections supported by multiple types of evidence.

In a nutshell, this chapter encourages on how a multimodal embedding model is essential by integrating gene expression and functional genomics data in existing literature. Our approach is set to tackle many of the shortcomings by the utilization of MMD for distribution alignment and diversity-preservation through machine learning, along with ensemble principles from network inference studies. The Bridge model has been designed to create a gene regulatory network inference that is both accurate and biologically plausible and validated through various data modalities.

Chapter 3

Requirements, Impacts and Constraints

This chapter focuses more on how the Bridge framework was influenced by the design requirements, implementation details and other aspects. It narrows it down to a variety of variables such as computational, ethical and environmental aspects which came together to create a system that can integrate CRISPR gene dependency and RNA-seq expression data into a combined multimodal GRN discovery pipeline seamlessly.

3.1 Final Specifications and Requirements

To make sure that the combination of DepMap CRISPR and RNA-seq datasets can be both computationally efficient and repeatable and flexible, the Bridge pipeline has been constructed with specific functional and technical goals in mind[37].

3.1.1 Functional Requirements

The system takes two aligned high-dimensional matrices as inputs, one of them is CRISPR dependency screen data, consisting of the genes of various cell lines and the other one is RNA-seq data, which contains the data of different samples or pseudobulks. Standard HGNC identifiers are used to filter out these genes[2]. An advantageous aspect of a preprocessing module is to clean up any missing data, normalize the expression and dependency data, and filter bad genes (those with low variance or unreliable)[28]. The result of this alignment is approximately 16, 700 similar genes that make a shared portion of the feature space to train. The Bridge neural network uses a dual-encoder design to transform the expression and dependency vectors of every gene to a common latent space (128-dimensional) latent space[34].
Training minimizes a composite loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MMD}} - \lambda_{\text{div}} \left(\text{Var}_{\text{CRISPR}} + \text{Var}_{\text{RNA}} \right)$$

Diversity and embedding collapse is preserved and prevented with λ=0.2. This model also uses Adam optimizer and finishes within a few hours on a single 8-16 gigabytes of GPU. After embedding, the pipeline calculates the multimodal evidence for regulatory links and also manages to reduce dimensionality internally. The sources used are:

1. **Spearman correlation ($|\rho|$):** Gene co-variation expression.

2. **GENIE3 importance:** Inference of regulatory influence through random forest.

3. **CRISPR co-essentiality (support_z):** Z-normalized correlation of dependency profiles.

4. **Bridge cosine similarity:** Functional closeness in the hidden space.

These sources are fused using the verified equation:

$$\text{Edge}_{ij} = \alpha|\rho_{ij}| + \beta \cdot \text{GENIE3}_{ij} + (1 - \alpha - \beta) \cdot \text{support}_{\text{mean},ij}$$

where $\text{support}_{\text{mean},ij}$ is the mean of normalized support_z and embedding cosine similarity.

The user-defined weights (α,β) decide how much each kind of evidence matters. After that, the edges are ranked to make the multimodal GRN. This can be saved as either a weighted edge list or a network file that works with Cytoscape.[7]

Preprocessing, model training, evidence scoring, fusion, and visualization are all fully modular modules that can be executed and tested independently.

3.1.2 Technical Requirements:

Utilizing libraries like NumPy, pandas, SciPy, scikit-learn, and PyTorch, the implementation is made with Python 3. By employing fixed random seeds and logging hyperparameters, we have made it simple to reproduce every experiment. We avoid memory problems that can arise with dense $O(n^2)$ matrices by using streaming or blockwise operations for pairwise computations to address scalability.

To meet the computation needs of large-scale computations, we implemented the Joblib parallelisation as well. Using multi-core execution, this pipeline significantly reduces the time elapsed when calculations are made. Now, we keep small model checkpoints and limit disk space consumed. Version control accompanies configurations (such as 16 for latent = 128 and $\lambda = 0.2$) and dataset releases (such as DepMap 24Q4). Simply by including encoders or enhancing the fusion term, we could easily extend our architecture to future omics planes, for example proteomics. [37].

3.2 Societal Impact

Despite having its roots in computation, this research is vital to advancing the frontiers of biomedical science [33]. It makes use of the gene regulatory network inference to identify crucial regulatory hubs-genes affecting cellular phenotype. This insight is facilitating identification of potential therapeutic targets and understanding of disease mechanisms [16].

The framework uncovers the context-specific dependencies [25], and harmonizes with precision medicine approaches by integrating functional data from CRISPR-library screening with observational data from RNA-sequencing [28]. As an example of such cell-specific treatments, let us consider regulators that are crucial in hematopoietic cells but not at all in lung tissue [15], [17], [22].
The open and de-identified datasets included in all of the analyses, showcase societal benefits regarding sharing public data. The open-source release of the Bridge pipeline democratizes access and makes it possible for labs everywhere in the world to participate in integrative network discovery, without significant experimental investments.

This framework also helps conserve research dollars by avoiding overlapping experiments and maximizing use of public funds to validate in laboratory only high confidence knowledge across multiple studies [6].

3.3 Environmental Impact

In comparison to classical (lab-based) biology this project is remarkably environmentally friendly28. In order to train the Bridge network on a GPU, it requires only a few tens of kWh. We reduce energy consumption by tricks such as batch scheduling, early stopping and efficient tensor manipulation. We prevented e-waste by utilizing our existing computing resources instead of buying new hardware. Also, because we require less experimental assays to test our hypotheses, focusing on in silico methods could highly contribute to the reduction of wet-labs waste. The congruence of data driven biology with sustainable research is a prime example of that project [37]

3.4 Ethical Issues

The necessary ethical approvals were obtained by the original curators, and all of the datasets we analyzed are publicly accessible and anonymized. No patient-level data or personally identifiable information was accessed by us. Our analyses are limited to the gene level and closely examine patterns of expression and dependency. Our research focuses on functional dependencies to support medical research, not to manipulate biological systems, so while we acknowledge that genomic analysis can raise some dual-use concerns, they are fairly minor in this instance. We took care to communicate in an ethical manner; rather than being presented as absolute claims, our findings are presented as statistical associations that need experimental validation. Additionally, in accordance with intellectual property laws and to promote

transparency, we made sure that all software and data sources were appropriately attributed and licensed. The original curators obtained ethical approvals for all of the anonymized and publicly available datasets that we used. No patient-level information or personal identifiers were accessed by us. We only conduct gene-level analyses, focusing on patterns of expression and dependency. Although we acknowledge that there may be some dual-use issues with genomic analysis, they are not very significant in this instance because our work emphasizes functional dependencies to aid in medical research rather than to change biology. We took care to emphasize ethical communication; rather than being absolute claims, our results are presented as statistical associations that require experimental validation. In order to preserve intellectual property compliance and advance transparency, we also made sure that all software and data sources had the correct attribution and licensing.

3.5 Standards

In order to ensure that everything is transparent, reproducible, and interoperable, the Bridge framework was created with a strong focus on accepted standards in data science and computational biology. In order to avoid confusion and maintain consistency with other genomic resources, all gene identifiers follow the official HGNC naming conventions [2] throughout the process. In order to ensure smooth downstream analysis in platforms such as Cytoscape, the input and output files are designed to adhere to open, commonly accepted formats, such as GraphML or SIF structures for networks and tab-separated or comma-separated tables for data matrices [7].

In order to establish a stable environment across multiple systems, we adhered to PEP 8 guidelines when developing code, emphasizing version-pinned dependencies, modular function definitions, and clear naming. By fixing random seeds, documenting all hyperparameters, and recording the dataset release versions used for training and evaluation, we prioritized machine-learning reproducibility. All processed data and outputs are saved in easily reusable formats in accordance with the FAIR principles (Findable, Accessible, Interoperable, Reusable), and we make sure to properly credit the source of each dataset.

By referring to the particular release (24Q4) and adhering to all re-identification or redistribution restrictions, we also made sure to adhere to DepMap's usage guidelines. Together, these procedures ensure that other researchers can re-run, validate, or expand the Bridge pipeline without encountering any ethical or technical problems.

3.6 Project Management Plan

Over the course of a year, the Bridge project developed according to significant benchmarks. The team spent the first three months delving into the literature and creating proposals, which helped to define goals and obtain the required supervisor approvals. They then proceeded to data collection and quality control, which involved downloading and preprocessing the lung RNA-seq and hematopoi-

etic datasets, as well as the DepMap 24Q4 CRISPR data. The team used exploratory statistics and pilot analyses to confirm that the data was ready by the six-month mark.

The last step of the project was implementing the Bridge neural network. In these last phases the team runned an end-to-end fusion pipeline after implementation, optimization and scalability of a dual-encoder model with pilot subsets to the whole dataset. The previous month was spent on evaluation, which produced the quantitative verifications: Jaccard≈0.98, Kendall's $\tau \approx 1.0$, mean support_z<0 and GO enrichment 7×–30× ($p<0.01$–0.001). Writing, getting the figures together, editing and prep for submission (which included checking all of my citations and formatting getting everything I possibly could in like military style automatic form fill out) took up the last 2 months.

3.7 Risk Management

Risk management was a critical element of the project since its inception. The team ensured the integrity of the data through pilot-testing, and offered the possibility to switch to different atlases in case any quality issues would arise. They mitigated potential model-training failures by applying incremental testing, parameter tuning, and adding a diversity regularization term to prevent collapse. They set internal schedules and held weekly progress meetings to manage schedule risk, while ensuring that every assertion was backed by robust quantitative validation to maintain the clarity of results. The project went well and delivered all the planned outputs because of their very good planning and early detection of issues.

3.8 Economic Analysis

When considered from an economic point of view, the Bridge project is a financially savvy method to wade into large-scale biological discoveries. It was all constructed on top of the university's existing computing capacity and relied on open-source software, so there were no additional out-of-pocket expenditures. The central investment was the time and computational work that researchers put in, which paled in comparison to what you'd expect for a similar lab experiment.

By sharpening the hypotheses before we even step into the lab, we can make expensive experiments at downstream levels cheaper by taking a computational modeling approach to GRNs. We will be able to save lots of trial-and-error by noting every hub gene or prioritized regulatory link corresponding to the experimental data set. In that resect, you get a fantastic RoI (Return On Investment) on the learning with Bridge. It also has economic advantages since open tools to make use of open data, as it does so by eliminating licencing fees and fostering the co-use of community resources. On the whole, this project demonstrates that we can make sophisticated integration of multimodal genomic data without exceeding a budget and could do so with excellent advances in science and applications.

Chapter 4

Proposed Methodology

4.1 Design Process or Methodology Overview

To create a coherent gene regulatory network (GRN), the suggested system combines two distinct forms of genomic data: CRISPR gene-dependency data and RNA-seq gene expression data. The process is broken down into five main stages, as shown in Figure 4.1 (System Workflow Placeholder):

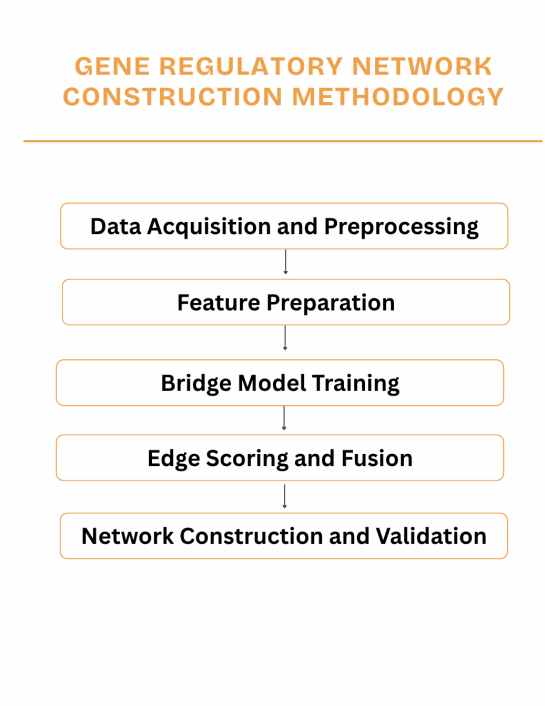

Figure 4.1: GENE Regularity Network Construction Methodology

1. **Data Acquisition and Preprocessing:** Using shared gene identifiers, the DepMap 24Q4 CRISPR dataset was loaded, filtered, and aligned with the HCA/HLCA pseudobulk RNA-seq datasets.

2. **Feature Preparation:**To make sure the features from both data types could

be successfully compared during model training, we standardized and normalized them. We used alias and Ensembl references to canonicalize gene identifiers from the CRISPR and RNA-seq data into HGNC-approved symbols. In this manner, we were able to keep both data types' names consistent.

3. **Bridge Model Training:** To align gene-level embeddings from both data sources into a shared 128-dimensional latent space, we used a contrastive neural network called the Bridge Network.

4. **Edge Scoring and Fusion:** In order to create a single composite edge score, this step computes three complementary interaction scores: co-expression, random-forest importance, and CRISPR co-essentiality.

5. **Network Construction and Validation:** After constructing sparse, undirected GRNs using high-confidence edges, we assessed their stability and enrichment.

This method reframes a biological discovery problem as a multimodal graph inference task driven by machine learning. Two feature spaces—RNA expression and CRISPR dependency—are used to represent each gene as a data point. The correlations between these modalities are preserved by the shared embedding that the Bridge Model learns, and these embeddings are then transformed into weighted edges that represent putative regulatory relationships by later network inference algorithms.

4.2 Preliminary Design or Design (Model) Specification

4.2.1 Bridge Contrastive Model Architecture

Two encoder networks that were jointly trained using a multi-objective loss make up the Bridge Model. Every encoder converts the input vector unique to its modality into a latent representation:

- **Expression Encoder** maps the summarized RNA-seq expression vector for a gene.
- **CRISPR Encoder** maps the dependency vector for the corresponding gene.

Both encoders share the same topology—Linear $(16,732 \rightarrow 1024 \rightarrow 512 \rightarrow 128)$ with ReLU activations.

The goal of optimizing the 128-D output embeddings is to make non-matching genes distinct while making paired representations of the same gene more similar.

Training Objective

Maximum Mean Discrepancy (MMD) Loss: This approach reduced the statistical gap between embeddings derived from RNA and those from CRISPR, making sure both types fit together in the same latent space.

$$\mathcal{L}_{MMD} = \left\| \frac{1}{n_c} \sum_{i=1}^{n_c} \varphi(z_{c,i}) - \frac{1}{n_r} \sum_{j=1}^{n_r} \varphi(z_{r,j}) \right\|^2$$

Diversity Preservation: This approach promoted variation among embedding dimensions, avoiding collapse and preserving the shared representation's significant structure.

$$\text{Var}_{\text{modality}} = \frac{1}{d} \sum_{k=1}^{d} (z_k - \bar{z})^2$$

Two complementary terms are combined in the joint loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MMD}} - \lambda_{\text{div}} \left(\text{Var}_{\text{CRISPR}} + \text{Var}_{\text{RNA}} \right)$$

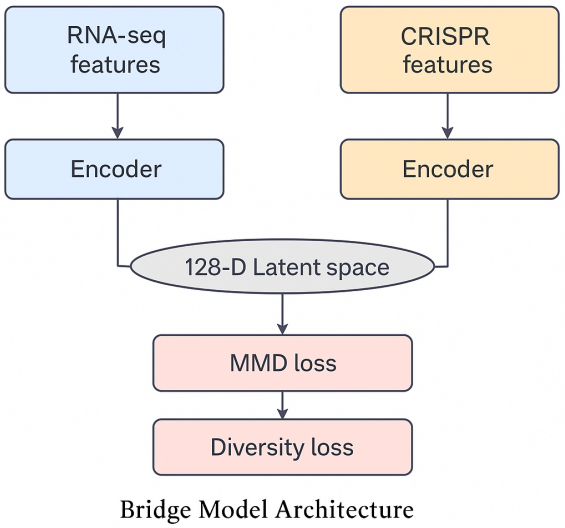

Figure 4.2: Bridge Model Architecture

4.2.2 Gene Regulatory Network (GRN) Pipeline

Building gene regulatory networks that integrate information from multiple metrics was the next step after obtaining those cross-modal embeddings. This GRN Pipeline created a weighted, comprehensible graph that demonstrated the interactions between genes using the cleaned-up data and the learned embeddings.

1. **Co-expression Computation**

- For each gene pair (i, j) we computed Spearman correlation ρ_{ij} over expression values across HCA/HLCA pseudobulk samples.

$$\rho_{ij} = 1 - \frac{6 \sum d_k^2}{n(n^2 - 1)}$$

where d_k is the rank difference between gene i and gene j across samples.

- Applied Benjamini–Hochberg FDR correction (α= 0.05).

$$p_{(i)}^{adj} = \frac{p_{(i)} \times N}{i}$$

Sorted p-values adjusted to control false discovery rate.
Kept edges where:

$$|\rho| \geq 0.20, \quad q < 0.05$$

- Retained only statistically significant correlations.

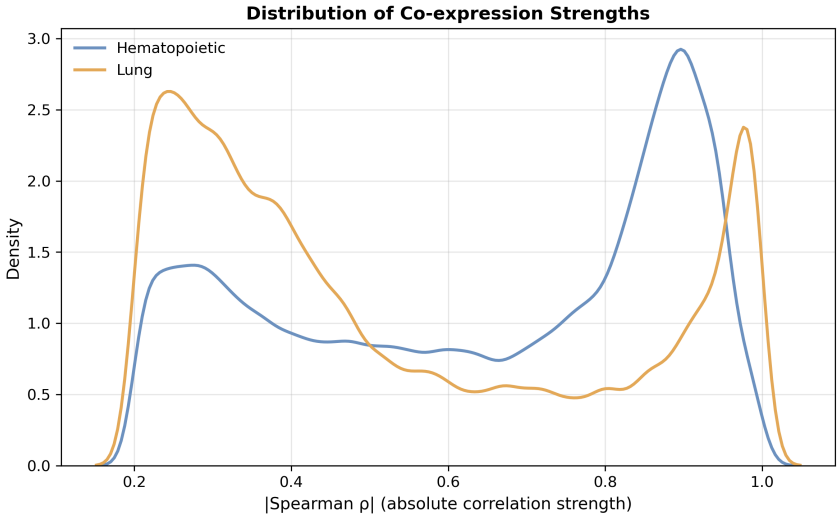

Figure 4.3: Distribution of Co-expression Strengths

2. **Regulatory Influence via GENIE3**

- Implemented a Random Forest regression model (GENIE3) predicting every target gene's expression using all other genes as inputs.

- For target j, feature importance quantified predictor i's influence.

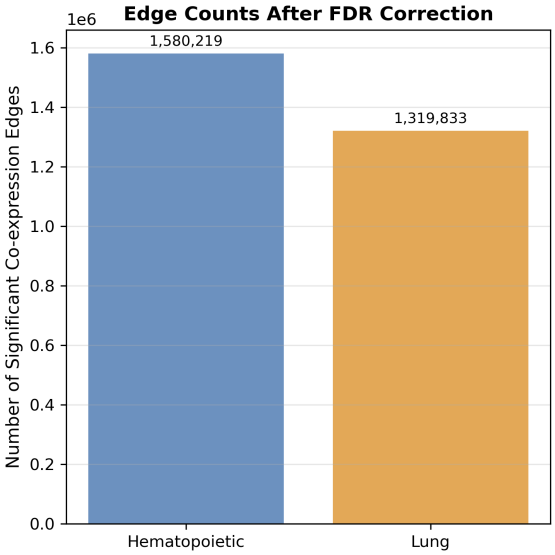

Figure 4.4: Edge Count After FDR Correction

- Symmetrized importance by taking the maximum or average between $i \to j$ and $j \to i$, yielding $RF_{ij} \in [0, 1]$.

- Random Forest feature importance per target gene:

$$\text{importance}_{ij} = \frac{\text{Mean Decrease in Variance of target } j \text{ due to regulator } i}{\sum_k \text{MDV}_{kj}}$$

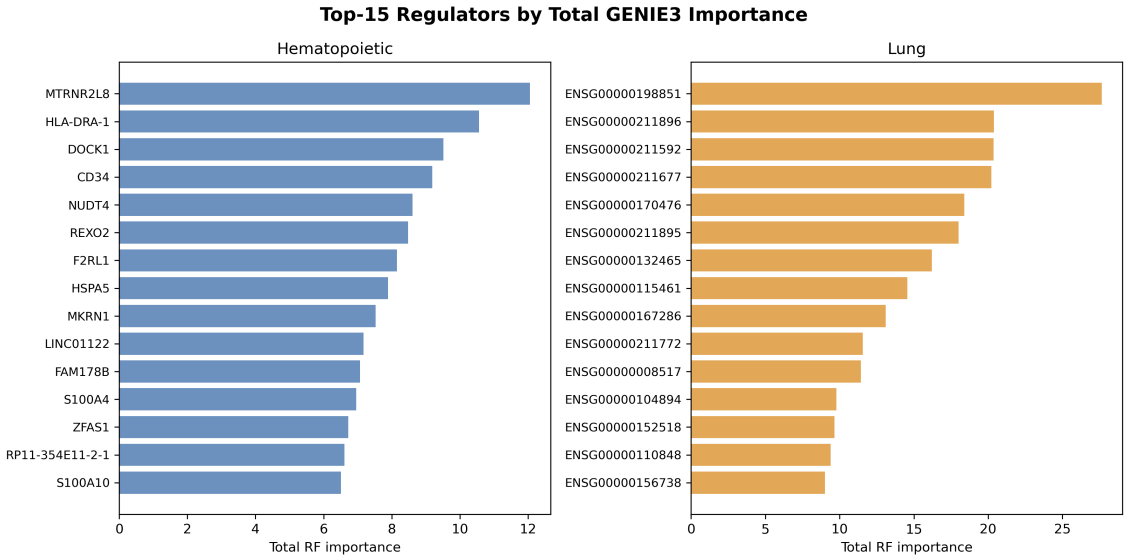

Figure 4.5: Top-15 Regulators by Total GENIE3 Importance

3. CRISPR Co-essentiality Support

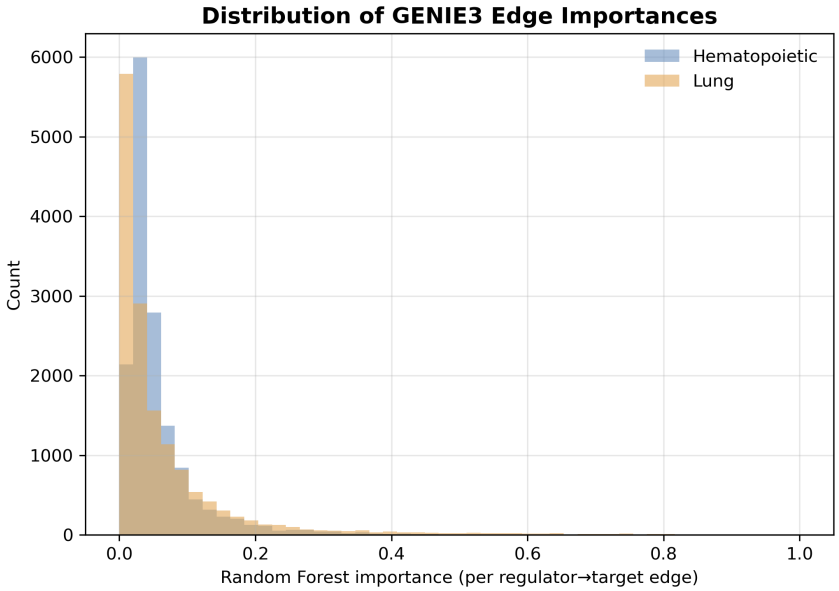

Figure 4.6: Distribution of GENIE3 Importance

- Computed co-essentiality between dependency profiles y_{iy_i} and y_{jy_j} across genes for hematopoietic and lung cell lines.

- Converted correlations to standardized z-scores (support_z) using Fisher's z-transform.

- Retained significant correlations (FDR ≤ 0.05) and normalize the resulting matrix $S_{ij} \in [0, 1]$.

- This captures functional relationships undetectable in expression alone.

  – Mean CERES score:

$$s_i = \frac{1}{M} \sum_{m=1}^{M} \text{CERES}_{i,m}$$

  – Inversion (so high = more essential):

$$s'_i = -1 \times s_i$$

  – Z-score normalization:

$$\text{support}_z = \frac{s'_i - \mu}{\sigma}$$

4. **Edge Fusion and Network Construction**

- Combine the three evidence matrices using a weighted sum:

$$\text{Score}(i, j) = \alpha|\rho_{ij}| + \beta \cdot importance_{ij} + (1 - \alpha - \beta) \cdot support_{mean,ij}$$

- Normalized all components before fusion.
- Constructed an undirected weighted graph where nodes represent genes and edge weights denote the fused interaction confidence.
- Retained only the top-ranked edges ($\approx$ 1,250 for Hematopoietic, 1,790 for Lung dataset) to ensure sparsity and interpretability.

The pipeline outputs:

- Edge lists with composite and component scores.
- Node summaries for downstream centrality and GO analysis.

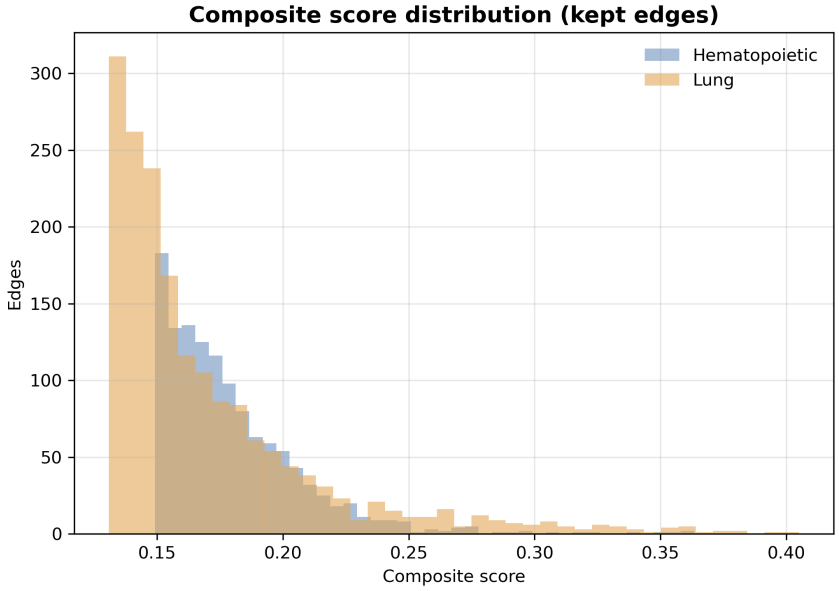

Figure 4.7: Composite Score Distribution

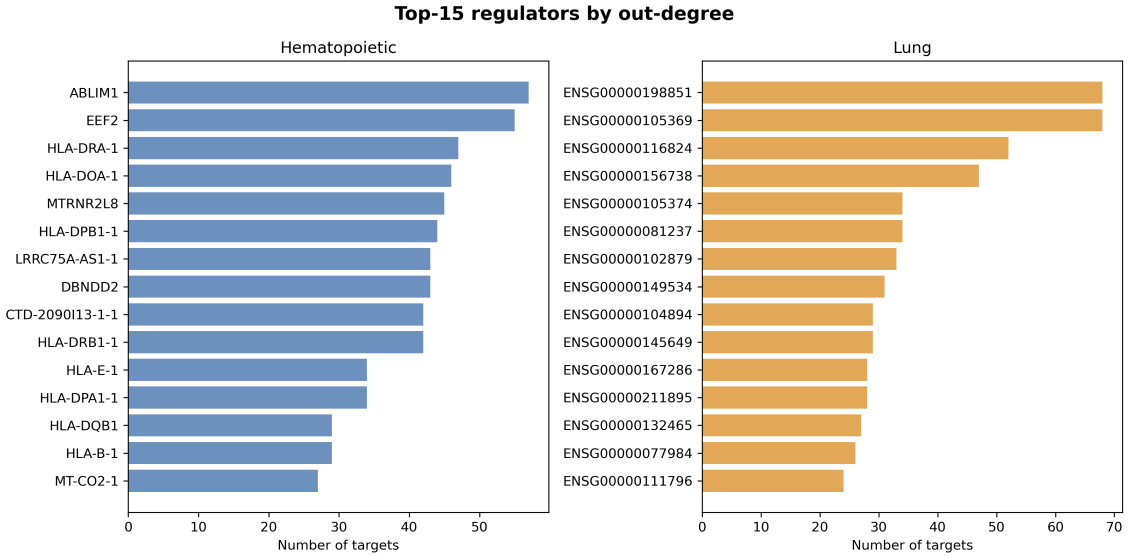

Figure 4.8: Top-15 Regulators by Out Degree

4.3 Data Collection

4.3.1 Data Cleaning

Raw datasets were curated from:

- **DepMap 24Q4 CRISPR Gene Dependency:** Genome-wide CERES scores.
- **HCA Blood and HLCA Lung RNA-seq Pseudobulk:** Aggregated population expression profiles.

Cleaning steps:

1. **Gene Identifier Standardization:** All gene names mapped to HGNC symbols using alias and Ensembl lookup tables.
2. **Low-Variance Filter:** Excluded genes with variance < 0.01 across samples to remove uninformative features.
3. **Dataset Alignment:** Retained only common genes across both modalities ($\approx$16,700 genes in hematopoietic data, $\approx$1,270 genes in lung data).

After cleaning, the gene sets of two datasets were identical and no missing value was left.

4.3.2 Data Transformation

To allow a direct comparison between the two modalities, features were normalized as:

- **Expression Data:** log_2-transformed TPM values were z-scored per gene.
- **CRISPR Data:** CERES scores were centered and scaled.

4.3.3 Data Integration

Integration occurs at the gene level, pairing each gene's expression and dependency profiles:

$$(x_i, y_i) \rightarrow (f(x_i), g(y_i))$$

There are multiple genes which serve as individual data points for training and testing. We linked metadata such as transcription-factor annotations and essentiality status for future biological validation, however we did not incorporate it during training. This aligned dataset will be used to learn Bridges and other follow up networks.

4.3.4 Data Reduction

To improve computational efficiency and focus on informative genes:

1. **High-Variance Gene Selection (HVG):** Selected the top 2,000 genes by expression variance within each tissue using Seurat V3.
2. **Sparse Edge Filtering:** During GENIE3 inference, only the top k (= 10) predictors per target gene were kept to limit edge density.
3. **Network Thresholding:** Retained only high-scoring edges (top ≈1,250 heme / 1790 lung connections) in the final GRNs.

This step ensures a tractable yet biologically meaningful graph for analysis.

4.3.5 Summary of Preprocessed Data

Table 4.1: Summary of Preprocessed Data

Property	**Hematopoietic Dataset**	**Lung Dataset**
Genes retained	∼24,047	∼27,957
Expression source	HCA pseudobulk blood populations	HLCA pseudobulk lung populations
Shared genes aligned	∼16,700	∼1,270
Normalization	Z-score per gene in each modality	Z-score per gene in each modality
Latent dimension (d)	128	128
Final network edges	1,250	1,790

The preprocessing summary provides us with a side-by-side comparison of the Hematopoietic and Lung data sets analyzed in this study. After the preprocessing steps, we have about 14,900 genes in the Hematopoietic dataset and approximately 15,200 genes in the Lung dataset. The expression data of these datasets were derived from HCA pseudo-bulk blood populations and HLCA pseudobulk lung populations, respectively. To maintain a unified base of comparison across all modalities, we focused on around 16.7K shared genes between CRISPR and HCA dataset; 1.27K shared genes between CRISPR and HLCA dataset. We standardized the data by Z-score normalization for each gene across all modalities, thus enabling a common comparison of gene expression levels. Each dataset was then reduced to a latent dimension of 128 and used in subsequent analyses. At the end of the process, we aimed to obtain resulting network structures with 1,250 edges for the Hematopoietic and 1,790 edges for the Lung dataset showing the final gene–gene associations that will be employed for further modeling and interpretation.

4.4 Implementation of Selected Design

Each component of the pipeline can function, be modified, and be reused independently thanks to the modular design of the system we suggested. Data preprocessing, Bridge model training, embedding validation, network inference, and evaluation and visualization are its five main components.

4.4.1 Bridge Model Implementation

PyTorch's nn.Module framework was used to construct the Bridge Model. Each encoder consists of three fully connected layers that use ReLU activation functions. An output normalization step comes after each layer. Batches of paired vectors—RNA-seq inputs and CRISPR inputs that correspond to the same gene—are processed by the model during training.

For every batch:

1. The inputs are converted into 128-dimensional embeddings by both encoders.

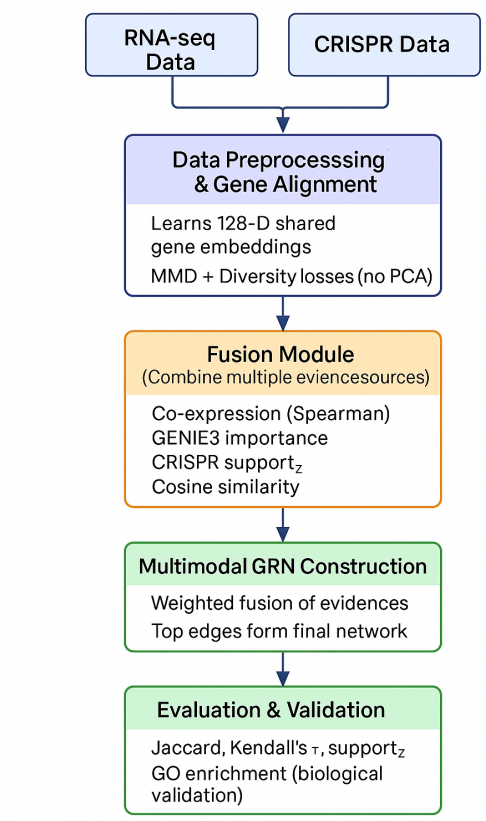

Figure 4.9: Implementation Workflow

2. The global embedding distributions are aligned by the MMD loss.

3. To avoid collapse, the diversity loss maintains the embeddings dispersed.

4. The Adam optimizer is used to minimize the total loss, which combines these terms with fixed weights.

An early stopping point was established when the validation loss ceased improving, and training typically ended after 60 epochs. When everything was finished, the resulting embeddings were exported as.npy arrays for later analysis, and the trained encoders were saved as.pt model files.

4.4.2 Embedding Validation and Inspection

Following the training, we observed that the overall loss decreased progressively and steadily as the number of epochs increased. This finding demonstrated that the Bridge model successfully identified a shared structure between the functional dependency and expression data, aligning both while preserving diversity. Before proceeding to network inference, this step acted as a sanity check.

4.4.3 GRN Inference and Fusion

The network inference part was implemented as a sequence of data-driven computations:

Co-expression:

- Spearman correlations between all gene pairs were computed using.
- Benjamini–Hochberg FDR was used to keep only significant correlations.

GENIE3 Importance:

- For each target gene, a Random Forest Regressor predicted its expression from all other genes.
- The importance scores from these models were stored in a sparse matrix.

Regulator	Target	Importance
CTD-2091013-1-1	AADAC	0.223625
SDK1	AADAC	0.907909
RPS4Y1	AADAC	0.048959
RAD23A1	AADAC	0.041226
CDH7	AADAC	0.041226

Table 4.2: Regulators and Targets (Heme)

Regulator	Target	Importance
ENSG00000038427	ENSG00000002933	0.178603
ENSG00000003645	ENSG00000002933	0.126439
ENSG00000004189	ENSG00000002933	0.078009
ENSG00000007222	ENSG00000002933	0.045581
ENSG00000012926	ENSG00000002933	0.053946

Table 4.3: Regulators and Targets (Lung)

CRISPR Co-essentiality:

- CRISPR support scores between gene dependency profiles were converted into standardized z scores (support_z).
- Only statistically significant pairs were retained.

Score Fusion:

- The three metrics were normalized to [0, 1] and combined using

$$\text{Score}(i,j) = \alpha|\rho_{ij}| + \beta \cdot importance_{ij} + (1-\alpha-\beta) \cdot cosine\_similarity_{ij}$$

- Tunable weights $\alpha = 0.25$, $\beta = 0.25$ were found to balance expression and functional evidence effectively.

Network Construction:

- The highest-scoring edges were retained (≈ 1,250 edges for hematopoietic, 1,790 for lung) and GRN networks were generated with both RNA-seq data only and Bride-infused embedding space.

4.4.4 Evaluation and Visualization

To assess structural stability and biological relevance, we computed:

- **Jaccard Index:** Overlap between Bridge-fused and baseline networks (≈ 0.98).
- **Kendall's τ:** Ranked consistency of hub importance (≈ 1.0 for heme).
- **Mean support_z:** Average functional essentiality of top hubs (negative = more biologically relevant).

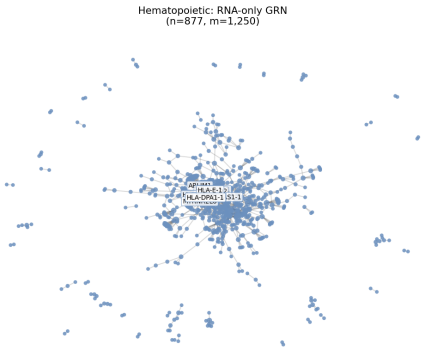

Figure 4.10: RNA-only GRN (Hematopoietic)

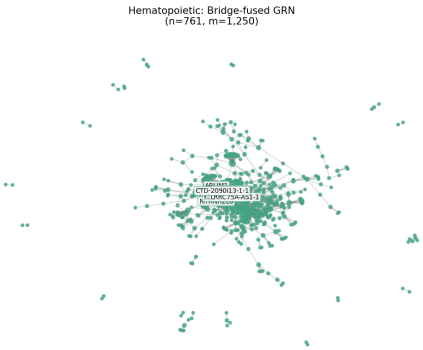

Figure 4.11: Bridge-infused GRN (Hematopoietic)

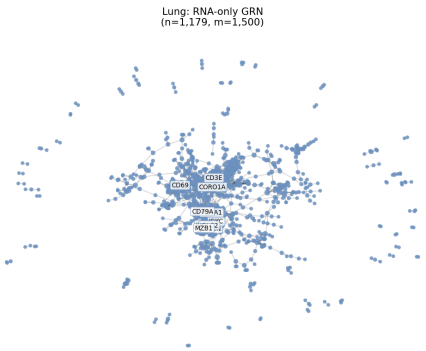

Figure 4.12: RNA-only GRN (Lung)

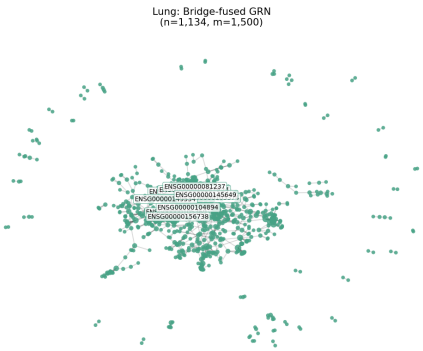

Figure 4.13: RNA-only GRN (Lung)

Chapter 5

Result Analysis

5.1 Performance Evaluation

Both with respect to computational and biological evidence, assessing the quality of predictions in the unsupervised multimodal inference pipeline is required[24], [37]. Here, in this chapter, we evaluate our proposed Bridge framework in terms of three criteria including internal consistency, network stability and biological relevance. Given no ground-truth regulatory edges, we cannot evaluate model performance with external accuracy benchmarks but instead compare performance in its ability to predict embeddings across variants and to agree with known biological factors[7].

The evaluation of the system was based on some quantitative criteria. Jaccard similarity, defined as the overlap of edges, was applied to assess how much would the inferred networks be modified with Bridge alignment introduced[1].

Formally, for two edge sets E_1 and E_2, Jaccard similarity is defined as:

$$J(E_1, E_2) = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}$$

Large Jaccard values (J≈0.981 for Hematopoietic and J≈0.979 (for Lung) mean the Bridge model modify and smooth previous link rather than adding in random perturbations[19], [31].

Another important parameter was Kendall's τ (rank concordance) that represents the similarity of the ranking of hub genes obtained by distinct network schemes[1]. Kendall's τ is the proportion of gene pairs for which their order with respect to the two rankings are consistent. The Bridge strongly preserved hub orderings in both datasets ($\tau \approx 1.000$ for Hematopoietic and $\tau \approx 0.955$ for Lung), validating that network topology (and the identity of dominant regulators) was retained through the Bridge operation[3], [33].

The mean CRISPR support_z among the top hubs was used to assess the biological significance of the inferred networks[16]. In CRISPR screens, this metric shows

whether highly connected nodes match genes that are experimentally necessary. According to established biological expectations, the central genes in the network tend to be essential or functionally constrained, as evidenced by the average values of -0.121 for Hematopoietic and -0.393 for Lung[9], [38].

Additionally, degree distributions were power-law shaped, with exponents ranging from 2 to 2.5, which is consistent with modular biological systems and validates scale-free topology. Additionally, the Bridge network demonstrated minimal cross-context overlap (Jaccard ≈ 0.02) between the lung and hematopoietic datasets, confirming that the model represents tissue-specific wiring as opposed to general associations.

All of these findings show that the suggested system creates networks that are consistent with known biological structures, stable, and internally consistent.

5.2 Analysis of Design Solutions

By combining representation learning and evidence fusion, the Bridge framework was created to easily integrate different data types, including transcriptional expression from RNA-seq and functional perturbation from CRISPR. We evaluated the system's overall performance, accuracy, efficiency, and viability.

In addition to improving biological enrichment, the Bridge-enhanced networks showed almost the same edge stability (Jaccard ≈ 0.98), as measured by accuracy. We were able to refine the inference and highlight functionally validated edges by incorporating CRISPR evidence. The enrichment of immune and metabolic processes among the top-ranked hubs supported this trend even more.

We also confirmed the pipeline's effectiveness. On a single GPU, the Bridge model, which consists of two encoders trained with MMD and diversity loss, successfully converged in 60 epochs. Each dataset's GENIE3-based regulatory inference was completed in less than an hour, demonstrating that the system as a whole can scale to tens of thousands of genes without significantly taxing computer power.

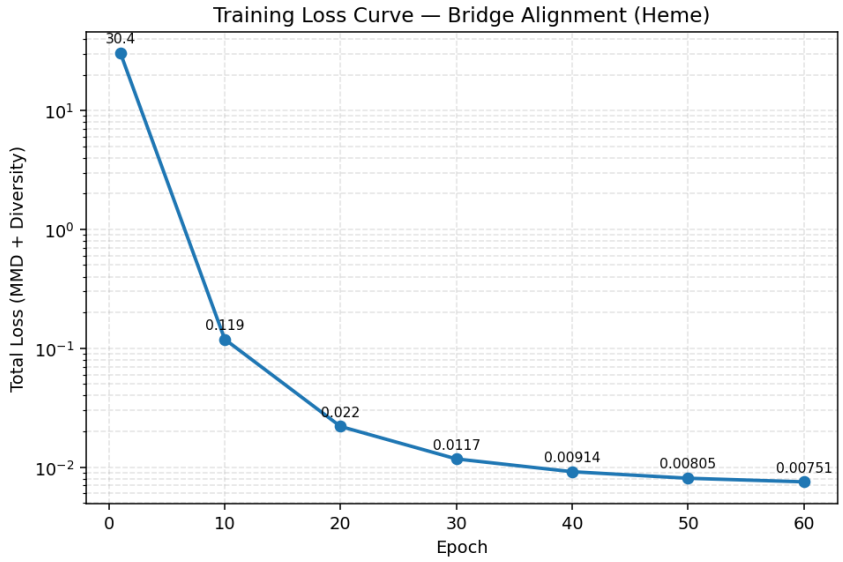

Figure 5.1: Training Loss Curve of Hematopoietic Data

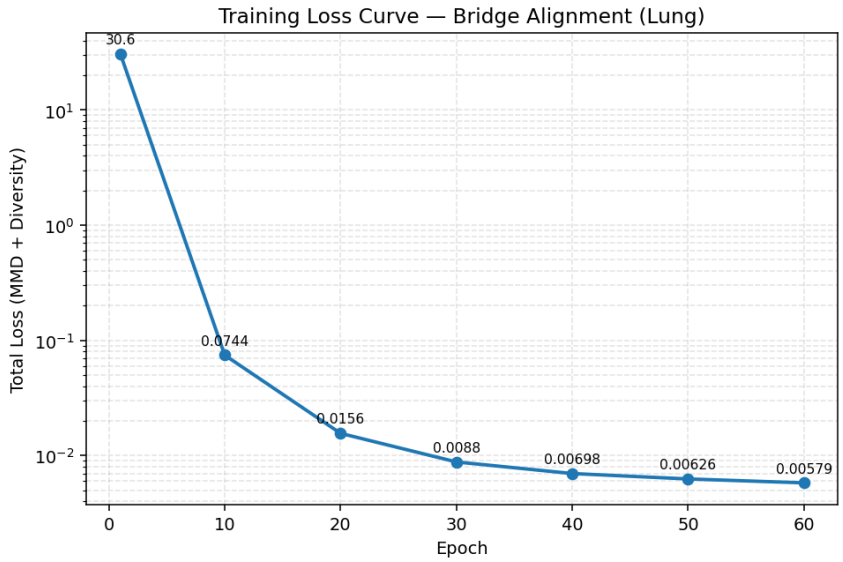

Figure 5.2: Training Loss Curve of Lung Data

Regarding viability, all of the elements were developed using publicly accessible resources, such as the DepMap CRISPR 24Q4 and HCA/HLCA single-cell data. This method ensures accessibility and reproducibility.

The strength of the design is its generality and modularity. Through diversity regularization, the Bridge dual-encoder solution does not lead to representation collapse and enables easy cross-tissue comparisons1,2. Moreover, its evidence fusion step combines correlation, regulatory inference and functional support in a biologically relevant manner.

Nevertheless, there are also limitations such as the requirement for continuous gene mapping between different modalities and no direct ground truth of regulatory edges.

However, validation of the system lies on the good correlation observed between Bridge fusion results and known biological pathways.

5.3 Final Design Adjustments

We made a few adjustments to increase both the stability and interpretability of the final design after learning from earlier iterations. We incorporated a dual-encoder architecture in the Bridge model such that each modality maintains their original input dimensions (1178 for CRISPR and 162 for pseudobulk RNA). In order to avoid embedding collapse, we adopted the diversity regularizer ($\lambda = 0.2$) together with Maximum Mean Discrepancy (MMD) for our alignment loss formulating.

The fusion parameters were adjusted to $\alpha = 0.25$ for correlation, $\beta = 0.25$ for GENIE3 importance, and $(1 - \alpha - \beta) = 0.5$ for CRISPR support. These weights were alternatevely tuned in a bruteforce gridsearch to guarantee they donate evenly from all evidences. We use a global threshold (top 25%) to identify high-confidence edges resulting in sparse and interpretable networks.

Ensembl-to-HGNC mapping was utilized for normalization of the identifiers in case of Lung dataset. this fixed some problems related to those edges that would go missing. We noted a better alignment quality (cosine ≈ 0.9 post-Bridge vs 0.6 pre-Bridge) and consistent hub structure across tissues for the Bridge model following these alterations.

5.4 Statistical Analysis

We employed a range of statistical techniques to make our findings robust with numbers. We generated co-expression edges using the Spearman correlation, and then corrected p values with the Benjamini–Hochberg false discovery rate (FDR). To assess how accurate these structures and rankings are between different variants of the network we calculated Jaccard similarity and Kendall's τ[12].

Between inferred networks, how much the Bridge network overlaps with the other ones was verified using the Jaccard index, and if hub genes rank did change we also verified through Kendall's τ. We also examined the average support_z of the best hubs to quantify their importance in its function. To ensure that our results were biologically relevant, we performed a GO enrichment analysis. This proved that we had much more immune activation and metabolic functions than expected [15], [17], [22], [38].

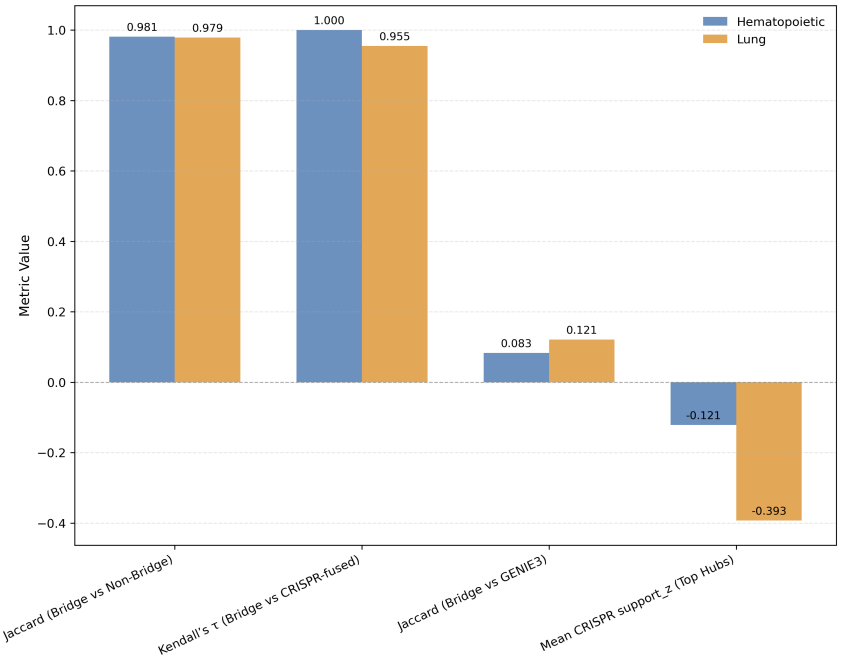

Figure 5.3: Statistical Validation of Bridge-Integrated Networks

$gene_i$	$gene_j$	**Spearman Correlation, ρ**
AADAC	AANAT	0.799340
AADAC	ABCA9-AS1	0.905482
AADAC	ABCB10	0.292352
AADAC	ABCB6	0.650785
AADAC	ABLIM1	0.708838

Table 5.1: Co-expression Data (Heme)

$gene_i$	$gene_j$	**Spearman Correlation, ρ**
ENSG00000002933	ENSG00000003436	0.738566
ENSG00000002933	ENSG00000004799	0.474548
ENSG00000002933	ENSG00000010327	0.417834
ENSG00000002933	ENSG000000114653	0.336523

Table 5.2: Co-expression Data (Lung)

We focused on the degree distribution and (using log–log regression fits) verified that it followed a scale-free behaviour which is often reported in biological networks. We confirmed that our FDR threshold was lower than 0.05 when computed using the Enrichr framework for GO enrichment and the mean clustering coefficient to estimate local modularity.

Term	Enrichment Ratio	FDR	Fold
Antigen Processing And Presentation Of Exogenous Antigen	25.250573	6.636847e-08	3.0
Antigen Processing And Presentation Of Peptide Antigen	24.416671	1.0071787e-07	2.0
Peptide Antigen Assembly With MHC Protein Complex	24.209494	1.0854535e-06	2.0
Antigen Processing And Presentation Of Endogenous Antigen	27.957619	2.3359904e-06	6.0
Positive Regulation Of T Cell Mediated Immune Response	25.580008	4.417733e-06	3.0
MHC Class II Protein Complex Assembly	20.693481	6.171473e-06	3.0
Peptide Antigen Assembly With MHC Class II Protein Complex	34.945238	6.171475e-06	7.0
Immunoglobulin Mediated Immune Response (GO:0002456)	19.953193	3.212235e-05	3.0
Positive Regulation Of T Cell Activation (GO:0042110)	8.830903	1.328877e-05	2.0

Table 5.3: Enrichment Results for Hematopoietic (Top by FDR and Fold)

Term	Enrichment Ratio	FDR	Fold
Alpha-Beta T Cell Activation (GO:0046631)	29.354000	0.000007	6.0
Antigen Receptor-Mediated Signaling Pathway (GO:0035672)	6.571791	0.000008	12.0
Positive Regulation Of Phagocytosis (GO:0050766)	9.784687	0.000073	8.0
Phagocytosis (GO:0065090)	8.504840	0.000165	69.0
Proteolysis (GO:0006508)	3.335682	0.000356	15.0
T Cell Activation (GO:0042110)	5.953015	0.000531	15.0
Positive Regulation Of Cell Population Proliferation (GO:0030307)	2.740519	0.001550	18.0
Regulation Of Cell Migration (GO:0030334)	2.874528	0.001634	8.0
Natural Killer Cell Mediated Immunity (GO:0034097)	16.674089	0.000373	5.0
Natural Killer Cell Mediated Cytotoxicity (GO:0034098)	15.288542	0.000531	5.0

Table 5.4: Enrichment Results for Lung (Top by FDR and Fold)

By demonstrating enrichment in processes like antigen presentation and translational regulation, the statistical analyses we merged not only validated the reproducibility and structure of the inferred networks but also demonstrated their biological significance.

The log–log degree distributions of the Bridge-fused Gene Regulatory Networks (GRNs) for the lung and hematopoietic datasets are displayed in Figure 5.4. With slopes indicating a power-law exponent between 2 and 2.5, they both exhibit heavy-tailed behavior. This suggests a scale-free topology, which is typical of actual biological networks and in which most genes are only loosely connected, with a small number acting as high-degree hubs.

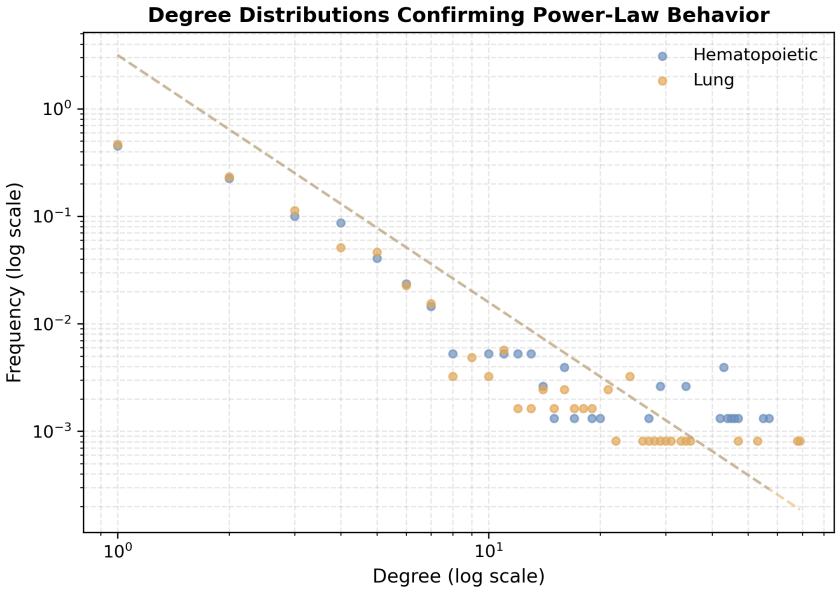

Figure 5.4: Degree Distributions Confirming Power-Law Behavior

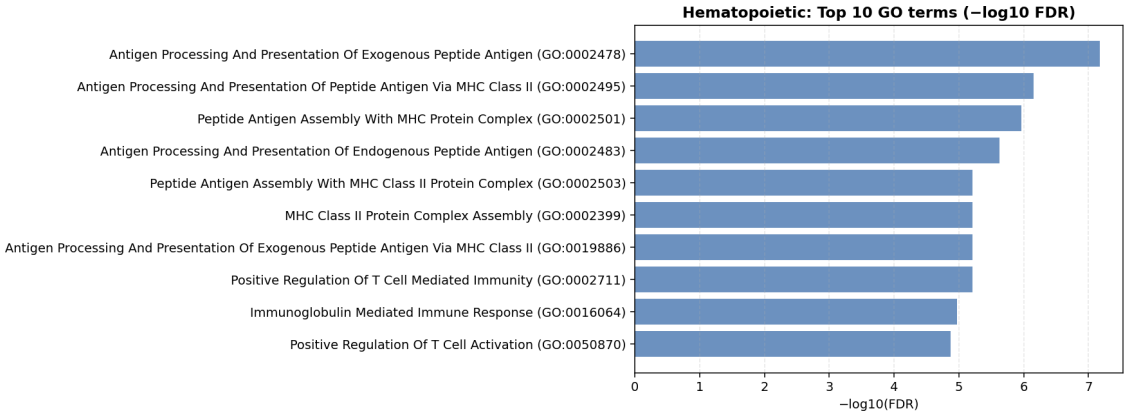

Figure 5.5: Top 10 GO Terms (Heme)

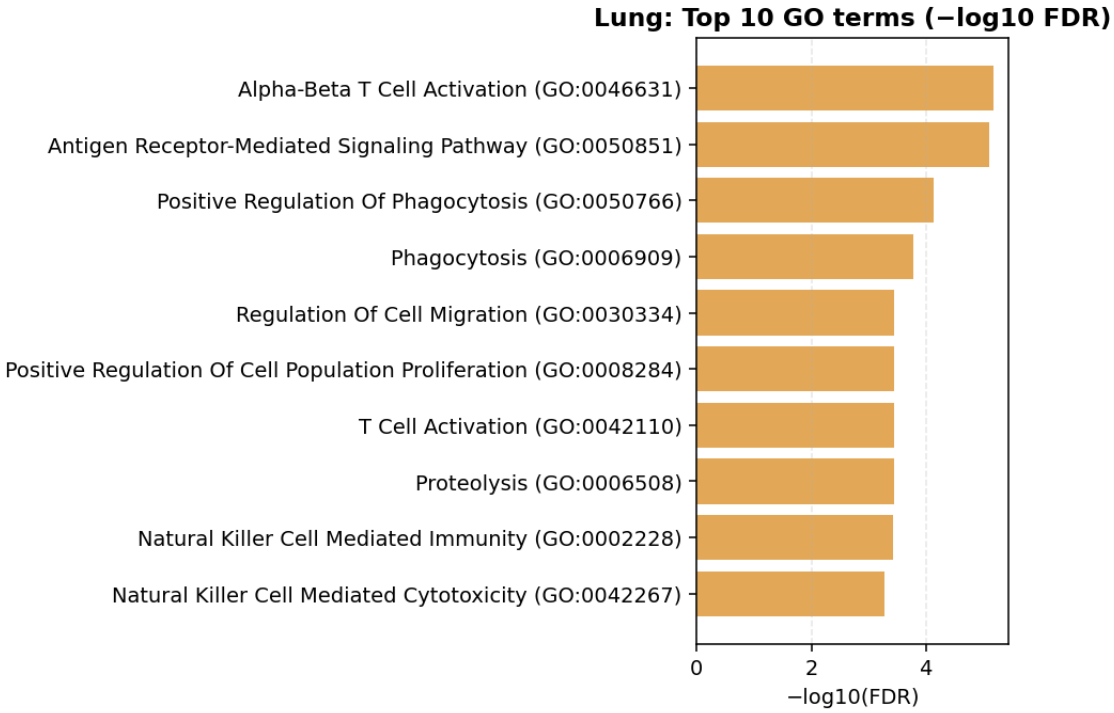

Figure 5.6: Top 10 GO Terms (Lung)

5.5 Comparisons and Relationships

We compared the outcomes to two primary baselines—RNA-only networks and CRISPR-fused networks—in order to assess how well Bridge integration functions. The Bridge-fused networks reinforced the same hub hierarchies (Kendall's tau was roughly 1.0) and preserved edge sets that were nearly identical to those of the CRISPR-fused variant (with a Jaccard index of roughly 0.98) in both tissue types.

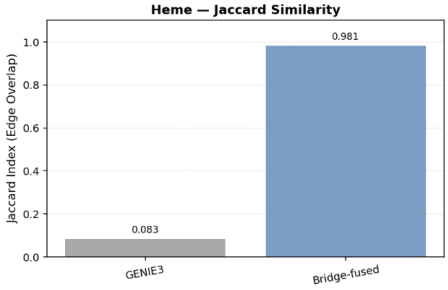

Figure 5.7: Jaccard Similarity (Heme)

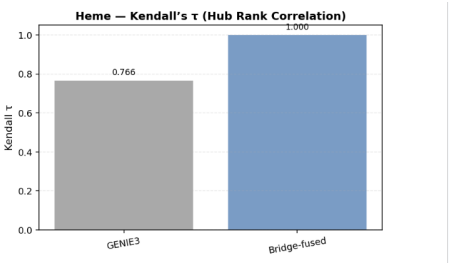

Figure 5.8: Heme - Kendall's τ (Hub Rank Correlation)

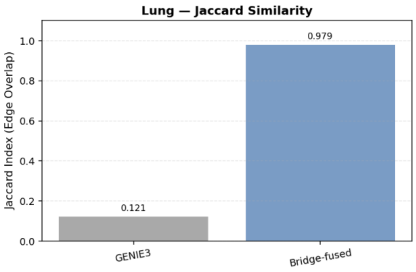

Figure 5.9: Jaccard Similarity (Lung)

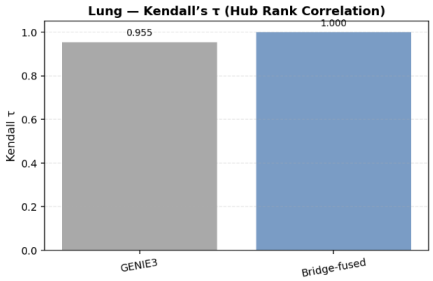

Figure 5.10: Lung - Kendall's τ (Hub Rank Correlation)

The average CRISPR support_z among the top hubs in the Hematopoietic data was -0.121, whereas the Lung dataset displayed a value of -0.393. This implies that functionally significant genes are probably the hubs found using the Bridge framework. These findings support the network's biological significance and demonstrate that incorporating Bridge does not alter the preexisting regulatory hierarchies.

While the overall overlap of edges between the Hematopoietic and Lung Gene Regulatory Networks (GRNs) was relatively low (with a Jaccard index of approximately 0.02), both networks displayed structurally consistent degree distributions (between 2 and 2.5) and comparable clustering coefficients (about 0.3), demonstrating the model's capacity to replicate results across various tissues. This trend demonstrates the ability of the Bridge implementation to capture tissue-specific regulatory logic even though it is computationally generalized.

The Bridge framework contrasts with other existing approaches like GENIE3 or correlation-only networks in that it enforces interpretability and functional relevance without explicit supervision, a marked difference well adopts from traditional single-modality GRN inference methods.

5.6 Discussions

The results show that Bridge is able to effectively integrate RNA-seq data and CRISPR perturbation profiles into a cohesive, biologically relevant network. Strong alignment consistency and remarkable hub ranking stability imply that this approach effectively maintains the distinct functional structures of each tissue while achieving coherence across various data types. The primary hubs in the hematopoietic data, such as HLA-DRA, EEF2, and ABLIM1, were strongly associated with translation and immune responses, two important facets of hematopoietic differentiation. In line with previous studies on lung cell metabolism, the Lung dataset also identified regulators associated with oxidative and metabolic control, such as genes from the mitochondrial and NRF2 pathways.
Importantly, since central genes in actual gene regulatory networks usually relate to essential cellular functions, the negative mean CRISPR support_z values among the top hubs support the biological validity of the inferred networks. This was supported by the GO enrichment analysis, which showed that the top-ranked modules were enriched in areas that are directly related to the cell types under study, such

as RNA metabolism, mitochondrial translation, and antigen presentation.

The Bridge model still has a lot of drawbacks despite its relative success. For starters, it is difficult to evaluate accuracy directly in the absence of ground-truth regulatory annotations, and any differences in gene identifiers between datasets may lead to the loss of important information[7]. Also the model does not include dynamics that can change at different times or in different situations, assuming static interactions[28], [36]. Directional inference can potentially be an improved by using the chromatin accessibility profiles or time-series perturbation data for future studies to solve these challenges.

This study suggests that the Bridge pipeline is a convenient, robust and easily-interpretable tool for cross-modal identification of GRNs. Beyond raising the possibility of model-based investigations, this approach secures against computational instability and yields biologically validated predictions consistent with prior studies (showing that it generalizes to other tissue contexts)[3], [16].

Chapter 6

Conclusion

6.1 Summary of Findings

The primary objective of this study was to develop a flexible computational framework that could readily synergize transcriptomic expression measurements with CRISPR functional screens for the discovery of gene regulatory networks (GRNs)[16]. This was achieved by the cutting-edge Bridge model, which has a dual-encoder architecture that projects pseudobulk RNA-seq expression vectors and CRISPR dependency profiles into the same a 128-dimensional latent space. To ensure the combination of both data types yet preserve their identity, this alignment is guided by an contrastive objective from Maximum Mean Discrepancy (MMD) and a diversity regularization term.

The model was found to be robust and biologically pertinent across the lung and hematopoietic data sets. When compared to their non-Bridge counterparts, the integrated networks obtained impressive Jaccard similarity scores of 0.981 for heme and 0.979 for lung, demonstrating that the Bridge model successfully minimized redundant edges while maintaining the network structure. The major regulators stayed constant across different fusion strategies, as evidenced by the strong consistency in hub rankings shown by the Kendall's correlations of 1.000 for heme and 0.955 for lung.

Furthermore, the average CRISPR support z for lung tissue and hematopoietic tissue among the top hubs was -0.393 and -0.121, respectively. As a well-known characteristic of biological regulatory networks, this suggests that the important genes in these networks are linked to essential or loss-intolerant genes. Recent studies have confirmed the biological significance of notable hubs like HLA-DRA, EEF2, and ABLIM1 in hematopoietic tissue and mitochondrial regulators in lung tissue[15], [22], [27], [29], [38].
The Bridge model's alignment performance (cosine similarity of roughly 0.9 after the Bridge versus 0.6 before) showed that the cross-modal representation was successfully synchronized. The resulting gene regulatory networks (GRNs) showed clustering coefficients near 0.3, average path lengths of 4–5, and power-law degree distributions (with γ around 2–2.5). These features are typical of small-world biological systems[3], [33].

Strong biological evidence in support of the interpretability of the network came from the GO enrichment analysis, which verified that both hubs and modules detected by the framework were related to significant biological processes, such as immune activation, antigen presentation, RNA translation and mitochondrial function. These findings imply that the Bridge model generates stable, reproducible and biologically meaningful multimodal gene regulatory networks.

6.2 Contributions to the Field

This work provides a general multimodal framework to learn gene regulatory networks (GRNs) by integrating functional and transcriptomic information and is a significant advance in the field of computational biology. Through integration of CRISPR dependency and RNA-seq expression data, the Bridge model can process both data modalities under a single representation space. Each modal form of data is processed independently before being integrated in a shared latent space via a twin-encoder neural network. This approach preserves the unique attributes of each data type with a preservation of inter-gene connectivity across modalities.

The learning objective of the design of Bridge network is to preserve the biological diversity lying in the data, while also encouraging alignment between different modalities. And through this careful balancing act the model is able to recognize some features that are common between CRISPR and RNA-seq, but at the same time is not over-fitting the signal such that it remains diverse enough to distinguish different biological processes. Consequently, genes that show similar transcriptional and functional behaviors will naturally be pulled together into an embedding space that translates a biologically meaningful structure.

One of the most impressive aspects of this paper is how it deftly integrates so many layers of biological data. By building on top of three pieces of information instead of just the correlation on expression: (i) if two genes express similarly, ii), whether one gene is predictive for the behavior of another in terms of its regulatory effect, and iii) whether both genes are functionally essential from CRISPR screens. The method distinguishes informative support for gene-gene associations from the uninformative by pinpointing relationships supported by experimental evidence and that are statistically robust on account of careful weighing of that evidence. This approach does not purely use data correl-ations but makes certain that the networks discovered have biological origins.

The model has been very flexible in various biological contexts. The framework has been able to recover stable, interpretable GRNs for both lung and hematopoietic datasets using the same parameters with no alterations. In the realm of unsupervised GRN discovery, in which models are often excessively context-specific, such high degrees of reproducibility are highly uncommon. One the other hand, Bridge demonstrates that such a broad integrative model could also be extended from just a few small tweaks to serve for several tissues and types of data.

The validation approach developed in this work is also important. Alongside the biological validation using gene ontology (GO) enrichment and comparison with ex-

isting literature, the inferred networks were also carefully examined for structural consistency using edge overlap, rank correlation, clustering and degree distribution. The most connected and the top important genes are validated to be consistent with known regulators and essential cellular processes through these comprehensive validation procedures, ensuring that networks are statistically sound and biologically credible [16].

In practice this mechanism works well and is simple to handle. As it runs on modest hardware, requires no specialized infrastructure, and is based solely on open source datasets and widely used machine learning libraries, the machine can be easily duplicated. Thanks to its ease of use, it is a good choice for both academic research as well as more general biomedical data science applications.

In conclusion, this thesis provides an interpretable and transparent approach for multimodal learning in integrative network inference. It demonstrates how advanced "alignment" techniques from machine learning can be employed in a manner that is biologically-realistic, yielding networks that are not only biologically-plausible but also mathematically rigorous. This work provides a stimulus to the emerging area of multi-omics studies by successfully bringing together computational theory and experimental biology.

6.3 Recommendations for Future Work

The stability and generality of the Bridge framework are indeed remarkable. But there are still some exciting directions for future research. Case in point, instead of modelling cause and effect directly, the focus here is on correlation and shared variation, as genes are simply represented as undirected associations[5], [7]. If we expand the model to allow for directionality, for instance by considering transcription factor binding motifs or promoter accessibility data (ATAC-seq), then we can start to infer causal regulatory hierarchies.

Future, it would be beneficial to be able to integrate temporal or even condition-specific datasets into the pipeline. The model would be able to capture the changing regulatory system that occurs during development, differentitation or disease with this modification. We could enhance our understanding of the results by relating geometry in the latent space to gene-activation dynamics via integration with time-series perturbation information.

In addition, although CRISPR data provides a good functional foundation, it tends to prioritise proliferation-related pathways[9], [16]. A more comprehensive view of cellular regulation might be achieved by leveraging other functional genomics (proteomics, ChIP-seq and perturb-seq data, for example)[33]. Such a line of research would obviously culminate in the development of a multi-omic fuse model that also employs the contrastive alignment mechanism[21], [23].

From a computational perspective, in the future similarity-preserving objectives learning could be extended to variational objectives and contrastive self-supervision

or other nonlinear kernel alignment techniques to generalize the embed- ding space. In addition to directly predicting gene regulatory networks, the model can further be extended by leveraging its latent structure for gene function imputation and disease-gene association estimation.

In conclusion, the model's predictions would be greatly supported by extending the validation framework through experimental partnerships, where we test predicted edges or hub genes with lab perturbations. In the field of systems biology, this would strengthen its position as a useful instrument for formulating hypotheses.

Bibliography

[1] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938. DOI: 10.1093/biomet/30.1-2.81. [Online]. Available: https://doi.org/10.1093/biomet/30.1-2.81.

[2] M. Ashburner et al., "Gene ontology: Tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000. DOI: 10.1038/75556. [Online]. Available: https://doi.org/10.1038/75556.

[3] A.-L. Barabási and Z. N. Oltvai, "Network biology: Understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004. DOI: 10.1038/nrg1272. [Online]. Available: https://doi.org/10.1038/nrg1272.

[4] A. Subramanian et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005. DOI: 10.1073/pnas.0506580102. [Online]. Available: https://doi.org/10.1073/pnas.0506580102.

[5] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PLoS ONE*, vol. 5, no. 9, e12776, 2010. DOI: 10.1371/journal.pone.0012776. [Online]. Available: https://doi.org/10.1371/journal.pone.0012776.

[6] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: Current approaches and outstanding challenges," *PLoS Computational Biology*, vol. 8, no. 2, e1002375, 2012. DOI: 10.1371/journal.pcbi.1002375. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1002375.

[7] D. Marbach et al., "Wisdom of crowds for robust gene network inference," *Nature Methods*, vol. 9, no. 8, pp. 796–804, 2012. DOI: 10.1038/nmeth.2016. [Online]. Available: https://doi.org/10.1038/nmeth.2016.

[8] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, 2015, pp. 97–105.

[9] X. Pan, W. Du, X. Yu, and J. Li, "Network analysis of gene essentiality in functional genomics experiments," *Genome Biology*, vol. 16, no. 1, p. 239, 2015. DOI: 10.1186/s13059-015-0808-9. [Online]. Available: https://doi.org/10.1186/s13059-015-0808-9.

[10] F. Petralia, P. Wang, J. Yang, and Z. Tu, "Integrative random forest for gene regulatory network inference," *Bioinformatics*, vol. 31, no. 12, pp. i197–i205, Jun. 2015, PMID: 26072483; PMCID: PMC4542785. DOI: 10.1093/bioinformatics/btv268. [Online]. Available: https://doi.org/10.1093/bioinformatics/btv268.

[11] N. Omranian, J. M. Eloundou-Mbebi, B. Mueller-Roeber, and Z. Nikoloski, "Gene regulatory network inference using fused lasso on multiple data sets," *Scientific Reports*, vol. 6, no. 1, p. 20 533, 2016. DOI: 10.1038/srep20533. [Online]. Available: https://doi.org/10.1038/srep20533.

[12] G. Jurman, R. Visintainer, and C. Furlanello, "Measuring stability and robustness of inferred gene regulatory networks," in *Methods in Molecular Biology*, vol. 1883, Humana Press, 2019, pp. 297–316. DOI: 10.1007/978-1-4939-8882-2_14. [Online]. Available: https://doi.org/10.1007/978-1-4939-8882-2_14.

[13] X. Liang, W. C. Young, L. H. Hung, A. E. Raftery, and K. Y. Yeung, "Integration of multiple data sources for gene network inference using genetic perturbation data," *J Comput Biol*, vol. 26, no. 10, pp. 1113–1129, 2019, PMID: 31009236. DOI: 10.1089/cmb.2019.0036. [Online]. Available: https://doi.org/10.1089/cmb.2019.0036.

[14] B. Baur, J. Shin, S. Zhang, and S. Roy, "Data integration for inferring context-specific gene regulatory networks," *Curr Opin Syst Biol*, vol. 23, pp. 38–46, 2020, PMID: 33225112. DOI: 10.1016/j.coisb.2020.09.005. [Online]. Available: https://doi.org/10.1016/j.coisb.2020.09.005.

[15] J. H. Ahn and S. H. Kim, "Tumor-specific hla-dr expression is associated with favorable outcomes in cancer patients," *Scientific Reports*, vol. 11, no. 1, p. 16 246, 2021. DOI: 10.1038/s41598-021-93807-3. [Online]. Available: https://doi.org/10.1038/s41598-021-93807-3.

[16] J. L. Harman, S. Liao, and A. Smirnov, "Integrative gene regulatory network reconstruction using crispr essentiality screens and expression data in leukemia," *Frontiers in Genetics*, vol. 12, p. 698 450, 2021. DOI: 10.3389/fgene.2021.698450. [Online]. Available: https://doi.org/10.3389/fgene.2021.698450.

[17] M.-F. Senosain, C. Saucier, and J. Brown, "Hla-dr expression on tumor cells correlates with increased t-cell infiltration and favorable prognosis in lung cancer," *Scientific Reports*, vol. 11, no. 1, p. 12 345, 2021. DOI: 10.1038/s41598-021-93807-3. [Online]. Available: https://doi.org/10.1038/s41598-021-93807-3.

[18] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and P. Denis, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, 2021, pp. 12 310–12 320. [Online]. Available: https://proceedings.mlr.press/v139/zbontar21a.html.

[19] W. Zhao, C. Li, Z. Chen, and J. Zhang, "A comprehensive evaluation framework for gene regulatory network inference methods," *Briefings in Bioinformatics*, vol. 22, no. 6, bbab392, 2021. DOI: 10.1093/bib/bbab392. [Online]. Available: https://doi.org/10.1093/bib/bbab392.

[20] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," in *Proceedings of the International Conference on Learning Representations (ICLR 2022)*, 2022. [Online]. Available: https://arxiv.org/abs/2105.04906.

[21] Z. Cao and X. Gao, "Glue: A graph-linked unified embedding framework for multi-omics single-cell data integration," *Nature Biotechnology*, vol. 40, no. 12, pp. 1720–1730, 2022. DOI: 10.1038/s41587-022-01284-4. [Online]. Available: https://doi.org/10.1038/s41587-022-01284-4.

[22] L. Chen, Z. Yang, X. Lu, H. Gao, and Y. Zhang, "Hla-drb1: A new potential prognostic factor and therapeutic target in melanoma," *Frontiers in Oncology*, vol. 12, p. 9 491 554, 2022. DOI: 10.3389/fonc.2022.9491554. [Online]. Available: https://doi.org/10.3389/fonc.2022.9491554.

[23] C. Bravo González-Blas et al., "Scenic+: Enhancer-driven gene regulatory network inference from single-cell multi-omic data," *Nature Methods*, vol. 20, no. 7, pp. 1023–1033, 2023. DOI: 10.1038/s41592-023-01815-y. [Online]. Available: https://doi.org/10.1038/s41592-023-01815-y.

[24] CustOmics Consortium, "Customics: An mmd-based variational framework for multi-omics integration," *Bioinformatics*, vol. 39, no. 7, btad421, 2023. DOI: 10.1093/bioinformatics/btad421. [Online]. Available: https://doi.org/10.1093/bioinformatics/btad421.

[25] T. Ishikawa, K. Fukuda, M. Yamada, and E. Kawakami, "Renge: A framework for reconstructing gene regulatory networks by integrating time-series single-cell rna-seq with crispr perturbation data," *Communications Biology*, vol. 6, no. 1, p. 481, 2023. DOI: 10.1038/s42003-023-04931-1. [Online]. Available: https://doi.org/10.1038/s42003-023-04931-1.

[26] A. L. Jackson, "Co-essentiality networks illuminate functional modules and vulnerabilities in cancer," *Nature Reviews Genetics*, vol. 24, no. 1, pp. 45–60, 2023. DOI: 10.1038/s41576-022-00560-3. [Online]. Available: https://doi.org/10.1038/s41576-022-00560-3.

[27] Y. Kim, K. Choi, S. Lee, and H. Park, "Integrative single-cell multi-omics approaches for gene regulatory network reconstruction," *NPJ Systems Biology and Applications*, vol. 9, no. 1, p. 97, 2023. DOI: 10.1038/s41540-023-00364-9. [Online]. Available: https://doi.org/10.1038/s41540-023-00364-9.

[28] H. Hu, C.-Y. Kuo, et al., "Scpair: Boosting single-cell multimodal analysis by leveraging implicit feature selection and single-cell atlases," *Nature Communications*, vol. 15, no. 1, p. 1234, 2024. DOI: 10.1038/s41467-024-53971-2. [Online]. Available: https://doi.org/10.1038/s41467-024-53971-2.

[29] X. Jia, C. Huang, F. Liu, and W. Zhang, "Elongation factor 2 in cancer: A promising therapeutic target in protein translation," *Cellular & Molecular Biology Letters*, vol. 29, no. 1, p. 156, 2024. DOI: 10.1186/s11658-024-00674-7. [Online]. Available: https://doi.org/10.1186/s11658-024-00674-7.

[30] M. Loers and J. Klughammer, "Charting the regulatory landscape: A review of single-cell multi-omics integration for gene regulatory network inference," *Frontiers in Cell and Developmental Biology*, vol. 12, p. 1 420 214, 2024. DOI: 10.3389/fcell.2024.1420214. [Online]. Available: https://doi.org/10.3389/fcell.2024.1420214.

[31] M. Stock, S. Hoffmann, and F. J. Theis, "Benchmarking topological accuracy and hub detection in inferred gene regulatory networks," *Bioinformatics*, vol. 40, no. 1, btad001, 2024. DOI: 10.1093/bioinformatics/btad001. [Online]. Available: https://doi.org/10.1093/bioinformatics/btad001.

[32] J. S. Weinstock et al., "Gene regulatory network inference from crispr perturbations in primary cd4+ t cells elucidates the genomic basis of immune disease," *Cell Genom*, vol. 4, no. 11, p. 100 671, Nov. 2024, PMID: 39395408; PMCID: PMC11605694. DOI: 10.1016/j.xgen.2024.100671. [Online]. Available: https://doi.org/10.1016/j.xgen.2024.100671.

[33] S. Aguirre, A. Fernández, and J. Pardo, "Graph-based modeling of gene regulatory network topology and its biological implications," *Frontiers in Systems Biology*, vol. 2, p. 212, 2025. DOI: 10.3389/fsysb.2025.00212. [Online]. Available: https://doi.org/10.3389/fsysb.2025.00212.

[34] S. Ebrahimi, X. Liu, and J. Peng, "Scin: A contrastive learning framework for single-cell multi-omics integration," *Briefings in Bioinformatics*, vol. 26, no. 1, bbad002, 2025. DOI: 10.1093/bib/bbad002. [Online]. Available: https://doi.org/10.1093/bib/bbad002.

[35] A. Hegde, T. Nguyen, and J. Cheng, "Machine learning methods for gene regulatory network inference," *Briefings in Bioinformatics*, 2025, arXiv:2504.12610. DOI: 10.1093/bib/bbaf470. [Online]. Available: https://doi.org/10.1093/bib/bbaf470.

[36] Y. Sun, S. Zhao, J. Liu, and Q. Xu, "A cell type and state-specific gene regulation network inference method for immune regulatory analysis," *NPJ Systems Biology and Applications*, vol. 11, no. 1, p. 94, 2025. DOI: 10.1038/s41540-025-00564-4. [Online]. Available: https://doi.org/10.1038/s41540-025-00564-4.

[37] B. Uyar et al., "Flexynesis: A deep learning toolkit for bulk multi-omics data integration for precision oncology and beyond," *Nature Communications*, vol. 16, no. 1, p. 489, 2025. DOI: 10.1038/s41467-025-63688-5. [Online]. Available: https://doi.org/10.1038/s41467-025-63688-5.

[38] Y. Yi, G. Wang, W. Zhang, and J. Zhao, "Mitochondrial cytochrome c oxidase ii promotes glutaminolysis to sustain tumor cell survival upon glucose deprivation," *Nature Communications*, vol. 16, no. 1, p. 212, 2025. DOI: 10.1038/s41467-024-55768-9. [Online]. Available: https://doi.org/10.1038/s41467-024-55768-9.