

COMPUTATIONAL METHODS FOR GENE REGULATORY NETWORK INFERENCE

A Dissertation
Presented to
The Academic Faculty

by

Zijun Wu

In Partial Fulfillment
of the Requirements for the Degree
of Doctor of Philosophy in the
Department of Biomedical Engineering

Georgia Institute of Technology/Emory University
December 2025

COPYRIGHT © 2025 BY ZIJUN WU

COMPUTATIONAL METHODS FOR GENE REGULATORY NETWORK INFERENCE

Approved by:

Dr. Saurabh Sinha, Advisor
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Gregory C. Gibson
School of Biological Sciences
Georgia Institute of Technology

Dr. Peng Qiu
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Ahmet F. Coşkun
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Hanjoong Jo
School of Biomedical Engineering
*Emory University/Georgia Institute of
Technology*

Date Approved: August 20, 2025

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Saurabh Sinha, for his invaluable guidance, support, and mentorship throughout my PhD journey. His insight, encouragement, and dedication have been instrumental in shaping both my research and my growth as a scientist.

I would also like to sincerely thank the members of my thesis committee — Dr. Hanjoong Jo, Dr. Peng Qiu, Dr. Greg Gibson, and Dr. Ahmet Coskun — for their time, thoughtful feedback, and support throughout this process. Their diverse expertise and constructive input have greatly enriched my research.

I am especially grateful to Christian Park and Dr. Hanjoong Jo for their collaboration and for generously providing the datasets that were critical to this work. Their contributions and willingness to share resources have made a significant impact on this dissertation.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
LIST OF TABLES	vi
LIST OF FIGURES	vii
SUMMARY	viii
CHAPTER 1. INTRODUCTION	1
1.1 Biological Significance of Gene Regulation and GRN	1
1.2 SPREd: A Simulator-Supervised Framework for GRN Reconstruction	2
1.3 Dissecting GRNs in Atherosclerosis with Single-Cell Multi-Omics	4
1.4 Summary	5
CHAPTER 2. SPREd: A simulation-supervised neural network tool for GRN reconstruction	6
2.1 Abstract	6
2.2 Introduction	7
2.2.1 Understanding Gene Regulation Network	7
2.2.2 Existing Approaches and Challenges in GRN Reconstruction	8
2.2.3 SPREd: A Supervised Solution Leveraging Simulated Data	10
2.3 Methods	11
2.3.1 Simulation of Expression Matrix from a GRN, using SERGIO	11
2.3.2 Performance Evaluation Metrics	12
2.3.3 Input Data Preparation in SPREd	12
2.3.4 Architecture of SPREd-ML	13
2.3.5 Architecture of SPREd-SP	13
2.3.6 SPREd Training Details	14
2.4 Results	16
2.4.1 SPREd: Supervised Learning Framework For GRN	16
2.4.2 SPREd Outperforms Leading GRN Inference Methods on Synthetic Benchmarks	20
2.4.3 Effect of Benchmark Diversity on Predictive Accuracy	24
2.4.4 SPREd Performance is Better than or Competitive with Leading Methods on Real Data	27
2.5 Discussion	30
CHAPTER 3. Single-cell transcriptomic and chromatin dynamics of the flow-induced reprogramming of endothelial cells in atherosclerosis	35
3.1 Introduction	35
3.2 Methods	38
3.2.1 Animal studies	38
3.2.2 Isolation of nuclei from carotid arteries for scATAC-seq assay	39
3.2.3 CRISPR interference TF validation	40

3.2.4	RNA Isolation, qPCR, and Bulk RNA-seq	41
3.2.5	Bulk RNA-seq Analysis	42
3.2.6	Analysis of differential gene expression between EC sub-types in scRNA-seq data	43
3.2.7	scATAC-seq processing and analysis pipeline	43
3.2.8	scRNA-seq and scATAC-seq integration	44
3.2.9	Chromatin co-accessibility and CRE identification	46
3.2.10	TF-CRE and gene relationships	46
3.2.11	Key TF identification in cis-regulatory analysis in EC subpopulation	47
3.2.12	Differential accessibility and enrichment analysis	47
3.2.13	Gene expression prediction using random forest regressor	48
3.2.14	Gene expression prediction using MLP	48
3.3	Results	49
3.3.1	Single-cell multi-omic analysis in atherogenesis	51
3.3.2	Reprogramming of endothelial cells by D-flow under hypercholesterolemia	53
3.3.3	Cis-regulatory analysis reveals transcription factors underlying endothelial cell reprogramming	54
3.3.4	Explainable machine learning framework for TF discovery in atherogenesis	58
3.3.5	Combining cis-regulatory and coexpression-based evidence provides a high-confidence set of TFs associated with endothelial cell reprogramming in atherogenesis	66
3.3.6	Functional Validation of TFs reveals regulation of pro-atherogenic gene programs	68
3.3.7	Global visualization of transcriptional regulatory profiles of genes	69
3.3.8	Neural network model combines co-expression and cis-regulatory evidence for GRN inference	72
3.4	Discussion	75
CHAPTER 4. CONCLUSION		79
APPENDIX A. SUPPLYMENTARY FIGURES		82
APPENDIX B. SUPPLYMENTARY TABLES		92
REFERENCES		97

LIST OF TABLES

- Table 2.1 Performance comparison of SPREd, PORTIA, ENNET, GENIE3 and TIGRESS. Average Precision (AP) (A) and AUROC (B) in evaluations on expression matrices with varying number of conditions (columns in expression matrix). Each performance metric (AP or AUROC) is calculated for individual target genes, and results averaged over 5000 genes from 50 GRNs (100 genes in each GRN). Expression data are simulated using GRNs with $n_{MR}=5$, $n_{TF}=100$, $d_{TF \rightarrow G} = 3-7$. Note: TIGRESS runs for 10 conditions did not complete successfully. 22
- Table 2.2 Mean AP (A) and AUROC (B) of six evaluated methods on benchmarks using the MacIsaac2, YEASTRACT type2 (YEASTRACT2), or YEASTRACT count3 (YEASTRACT3) gold standard networks and the Nat Var (NV), Knockout (KO), and StressResp (SR) expression data from Siahpirani & Roy. AP/AUROC values are calculated for individual genes and averaged over genes. Highest value in each column is shown in bold. Median value across the nine benchmarks is shown for each method in the rightmost column. 28

LIST OF FIGURES

- Table 2.1 Figure 2.1 SPRED: a simulation-supervised learning framework for gene regulatory network (GRN) inference. (A) Standard approaches typically build ML models of the target genes using the expression levels of TFs as features. GRNs are then constructed based on the feature importance of TFs (features) in the model trained for a target gene. In SPREd, an ML model is trained to directly predict TFs regulating a target gene, based on expression matrix of all TFs and the target gene. The ML model is trained on simulated expression matrix-GRN pairs and can then be used to predict the GRN for any expression matrix. (B) Architecture of SPRED-SP neural network model. Given an expression matrix whose rows represent n_{TF} TFs and one target gene (panel A), the preprocessing step creates five matrices of pairwise relations (features) for each TF-TF pair and each TF-target gene pair. These features include covariance, Pearson correlation, Spearman correlation, mutual information and precision matrix entry corresponding to the TF-TF or TF-target gene pair. The five features of every gene pair involving a particular TF (say TF_i) then serve as the inputs of a 1D convolutional neural network (CNN) (input feature matrix, size $(n_{TF}+1) \times 5$). The feature map resulting from the first layer of convolution (out channel = 16) is of dimension $(n_{TF}+1) \times 16$, and feeds into a second convolution layer, whose outputs are fully connected to a hidden layer, which finally connect to the output layer. The output layer consists of a binary label indicating if TF_i is a regulator of the target gene (details shown in Methods). 15
- Figure 2.2 Figure 2.2 SPRED exhibits superior performance on synthetic datasets. (A) Schematic of synthetic data generation. Each synthetic data set comprises a GRN (left) and an expression matrix (bottom right). The GRN has three “layers” – master regulators (MR), transcription factors (TF) and target genes, with regulatory edges from one layer to the next. The MRs are included as the first layer so as to induce co-expression among TFs, mimicking real data. Parameters describing the GRN include the number of MRs (n_{MR}), the number of transcription factors (n_{TF}), the number of target genes (n_G), the number of incoming edges to each target gene ($d_{TF \rightarrow G}$), and the number of incoming edges to each TF ($d_{MR \rightarrow TF}$). A GRN is sampled at random while respecting these parameters and is used by SERGIO (middle), a biophysics-based model, to simulate the expression profiles of different artificial biological conditions, each of which is described by the production rates of MRs (top 19

right), thus generating an expression matrix whose rows include target genes, TFs and MRs and columns represent biological conditions. (B) Co-expression statistics of synthetic expression data. Absolute value of Pearson correlation coefficient (PCC) of TF-gene pairs (left) that comprise GRN edges (“true edges”) and those that do not (“false edges”) and of TF-TF pairs (right). Results are from simulations using GRNs in default configuration. (C,D) Average precision or “AP” (C) and AUROC (D) of the six evaluated methods – SPREd-SP, SPREd-ML, PORTIA, ENNET, GENIE3 and TIGRESS – on data sets with varying numbers of conditions (columns in expression matrix). Each performance metric (AP or AUROC) is calculated for individual target genes, and results summarized over 5000 genes from 50 GRNs (100 genes in each GRN). (E) Direct comparison of average precision (AP) between SPRED and PORTIA (top) or TIGRESS (bottom), for expression data with 50 conditions. (F) AP of SPREd-SP when using all but one (left) or only one (right) of the five features describing each TF-gene or TF-TF pair. AP when using all five features is shown by blue dashed line.

- Figure 2.3 Effect of benchmark parameters on GRN reconstruction. Performance (average precision) of SPREd-SP with varying edge density ($d_{TF \rightarrow G}$) of 1-2, 3-7, and 8-10 TFs per target gene (A), varying numbers of MRs (n_{MR}) (B), varying numbers of TFs (n_{TF}) (C), and varying levels of dropout added to the synthetic expression matrix (D). 25
- Figure 2.4 Performance comparison on heterogeneous benchmarks. AP comparison of SPREd-SP, SPREd-ML, PORTIA, ENNET, GENIE3 and TIGRESS on heterogeneous datasets comprising GRNs with (A) $n_{MR} = 5, 10$ and 40 (in equal numbers) and (B) with $d_{TF \rightarrow G} = 1, 2, \dots, 10$ (in equal numbers). 27
- Figure 3.1 Single-cell multi-omics overview. 50
- Figure 3.2 Cis-regulatory analysis of differentially expressed genes (DEGs) between endothelial cell (EC) sub-types associated with atherogenesis. (A) Illustration of how associations between TFs and DEG sets are discovered. Venn diagram showing the overlap between a DEG set (EC5-EC1) and the target gene of a TF (ATF3) reveals a substantial overlap, which is found to be statistically significant (p-value < 0.001 for motif-based analysis and p-value < 0.0001 for ChIP-based analysis) by a Hypergeometric test. On the right is shown the null hypergeometric distribution and the value of the observed overlap size (red arrow) is clearly significant. (B, C) 56

Most significant TFs associated with each DEG set based on motif analysis (B) or ChIP analysis (C).

Figure 3.3 Expression-based discovery of TF-gene regulatory relationships. (A) Schematic of computational workflow. The input comprises scRNA-seq data, shown here as a matrix with rows being genes and columns representing cells. The input also includes cell type annotations of cells and a subset of genes (rows) marked as TFs (candidate regulators). The expression data are subjected to normalization and imputation (using MAGIC) to obtain a processed expression matrix for further analysis, where each non-TF gene is modeled as a function of all TF genes using either Random Forest (RF) or MultiLayer Perceptron (MLP) regression. The cell type is also used as a covariate in this ML model. The trained ML model is interpreted by using “Integrated Gradients” (IG) scores that assign an importance value to each TF, reflecting the TF’s regulatory influence on the gene, which yields the gene regulatory network (GRN) edges leading into the gene. This is repeated for each non-TF gene. TFs with the largest number of target genes identified in this manner are considered the “important TFs”. The MLP architecture used is illustrated in the panel on top right. The input layer includes the expression level of all (n_{TF}) TFs and cell type (CT) of a cell, and the output is the target gene’s expression in that cell. Separate models are trained for each target gene. (B-D) Evaluation of ML models for predicting gene expression. Test R² values are shown for a collection of XX target genes (DEGs of EC1 vs EC2-5), comparing (B) MLP model to Random Forest, (C) MLP model with a 90% dropout rate with or without cell type (CT) as a covariate in addition to all TFs, (D) MLP model with a 90% dropout rate in the first fully connected layer versus without dropout. (E) Important TFs for each DEG set. Shown is the number of target genes of a TF among genes of each DEG set, as determined from the GRN inference outlined in panel A. Top 10 TFs from each DEG set were chosen based on the top rank of the number of regulating edges and then were concatenated across DEG sets.

61

Figure 3.4 (A) Top TFs derived for each DEG set obtained by combining evidence from cis-regulatory analysis and expression-based analysis. Each TF is marked (blue cells) as important to a DEG set if the TF-DEG set association is statistically significant under both motif-based and ChIP-based analyses and expression-based analysis finds the TF to be among the top 10 by number of targets in the DEG set. TFs are sorted by the row sum of the one-hot matrix then sorted by difference between the sum of EC1_EC* group and EC*_EC1 group. (B) Visualization of expression covariation between TFs and their predicted target genes, through six selected

65

examples. Each row of panels represents a TF (one of the top six rows panel A) and one of its predicted targets (prediction supported by motif-based and ChIP-based analysis as well as expression-based analysis). The TF name is indicated on the left and the target gene name is on the top. Each panel is a 2D histogram of the TF's and target gene's joint expression distribution across a population of cells. Axes represent expression values scaled to a 0-10 range for each gene or TF. In each row, the left-most panel shows the 2D histogram across cells of all cell types, while the second, third and fourth panels show the histograms for endothelial cells (EC), smooth muscle cells (SMC) and macrophages respectively, the three most common cell types in the data. These four panels use the same "bins" of expression values. The fifth panel ("equalfreqEC") shows the same histogram as the second, i.e., for endothelial cells, but the "bins" of expression are set to be deciles of expression among endothelial cells, for easier visualization of TF-gene expression co-variation in this sub-population of cells.

Figure 3.5 Experiment validation of top TFs predicted by cis-regulatory and expression-based analysis. (A) Results of qPCR validating successful CRISPRi-mediated knockdown of CREB3L2, MEF2C, and RELA, three of the top six TFs identified in the integrative analysis (Figure 3.4A), compared to the non-targeting control. (B) Volcano plots showing differentially expressed genes from bulkRNAseq comparing the CREB3L2, MEF2C, and RELA CRISPRi-knockdown cells and non-targeting control cells. Colors indicate if a gene is significantly DE at p -value < 0.05 and/or has a \log_2 fold-change $> 1e-6$ in absolute value, or neither (NS). (C) Heatmap on the left shows the log fold-change in expression upon knockdown of TF compared to control cells, for key groups of genes identified previously. Proatherogenic marker genes were significantly down-regulated after knockdown of CREB3L2, MEF2C, and RELA, while such strong down-regulation was not observed for EC functional and proatherogenic FIRE marker genes. (Warm colors represent down-regulation on TF knockdown.) Heatmap on the right shows log fold-change in expression of the same genes between the EC1 subtype and each of the more proatherogenic sub-types (EC2, ... EC5). (Warm colors represent up-regulation in the EC1 subtype.)

70

Figure 3.6 Enhanced prediction of gene regulation and classification of DEG sets through modality fusion. (A) Schematic of a classifier for EC DEG sets. The architecture of the neural network-based classifier was designed to predict a gene's membership in each of eight DEG sets (DE between EC1 sub-type and other EC sub-types). The classifier takes results from motif-, ChIP-seq, and expression-based analysis as input and learns to categorize the input into EC1_EC*

73

and EC*_EC1. The input feature vector has three dimensions (one for each type of evidence) for each TF, for a total of 3 x #TFs dimensions. (B) t-SNE plots showing embeddings of each gene obtained from the last hidden layer of the neural network of (A). Each panel shows the genes of a DEG set (EC1_EC*, EC*_EC1) in blue. Each vertically aligned pair of panels shows genes up- (top) and down- (bottom) regulated in EC1 compared to another EC subtype. (C) Comparison of accuracy of the baseline MLP model with uniform dropout and the modified MLP (fused model) that uses cis-regulatory evidence to inform dropout rates for the input layer. Shown are the test R2 values for 0.96 randomly selected genes, representing accuracy of predicting gene expression across cells in an unseen random subset of cells, based on TF expression profiles of those cells. (D) Scatter plot that compares occurrence of top-ranked important TFs. Each data point represented a TF with its count of as the top 10 regulators of a target gene in weighted dropout model (y-axis) and uniform dropout model (x-axis).

SUMMARY

Gene regulatory networks (GRNs) are commonly used to describe the complex regulatory relationships of transcription factors (TFs) and target genes. These networks are essential blueprints for a vast array of biological processes, including cellular development, response to stimuli, and disease progression. Despite their importance, accurate inference of GRNs from high-throughput data remains a fundamental challenge in bioinformatics due to dynamic and context-specific nature of cellular programs and the inherent noise in multi-modality, high-dimensional biological data.

This thesis presents a combination of computational approaches for GRN inference that address different aspects of these challenges. In the first project, we develop a novel simulator-supervised neural network framework, SPREd, for GRN reconstruction from transcriptomic data. This algorithm development project leverages synthetic data generation through a biophysics-based simulation model, allowing for training deep neural networks that directly predict the relationships between TFs and target genes. The approach we develop here offers an alternative to the established paradigm of training multi-variable models to predict a gene's expression from the levels of TFs. Our simulator-supervised learning strategy addresses the common limitation of insufficient ground truth data in GRN inference by creating large-scale synthetic datasets that capture realistic regulatory dynamics. We test SPREd on diverse synthetic and real data sets, demonstrating its improved accuracy in identifying both direct and indirect regulatory relationships compared to other GRN inference models.

While the SPREd model enables us to learn the underlying regulatory logic from bulk gene expression profiles alone, the second project in this thesis focuses on the reconstruction of GRNs from single-cell multi-omics data in disease contexts. Specifically, we aim to identify key regulators driving atherogenesis through integrative analysis of scRNA-seq and scATAC-seq data from multiple time points and experimental conditions. We use a combination of cis-regulatory analysis and coexpression-based GRN inference to identify TFs that regulate atherogenesis-linked transcriptomic shifts in endothelial cells. We also develop a new method that incorporates cis-regulatory evidence as prior information in a neural network model of gene expression, which subsequently yields GRN edges through explainable machine learning techniques. We present a novel strategy for visualization of regulatory profiles of genes, by using supervised learning to map each gene to a low dimensional embedding, allowing a global view of all genes that captures their differential expression patterns and regulatory evidence.

Our systematic analysis provides insights into the regulatory programs controlling flow-induced reprogramming of endothelial cells during the development of atherosclerosis. We predict the TFs *Creb3l2*, *Rela*, and *Mef2c* to coordinate the transition from proatherogenic endothelial cells to pathological states, and test these predictions in vitro through collaboration.

Together, this thesis demonstrates that combining principled computational innovation with biological applications can overcome fundamental limitations in GRN inference, establishing a framework for more accurate regulatory network reconstruction and mechanistic understanding of disease processes.

CHAPTER 1. INTRODUCTION

1.1 Biological Significance of Gene Regulation and GRN

Gene regulation is one of the most fundamental processes in biology. The precise spatiotemporal control of gene expression is essential for virtually every aspect of cellular function, from the basic metabolic processes that sustain life to the complex developmental programs that transform an embryo into a multicellular organism with hundreds of distinct cell types (Levine & Tjian, 2003; Spitz & Furlong, 2012). The importance of gene regulation extends far beyond normal cellular function: dysregulation of gene expression underlies the pathogenesis of many diseases, including cancer (Bradner et al., 2017), neurological disorders (Qureshi & Mehler, 2012), and cardiovascular diseases (Libby et al., 2009) as the leading causes of death. Understanding the mechanisms of gene expression is therefore crucial for not only advancing our fundamental knowledge of biology but also developing therapeutic strategies for human disease.

Gene regulation involves a combination of components such as transcription factors (TFs), splicing factors, long non-coding RNAs, microRNAs, and metabolites. Regulation of transcription, effected by TFs and their DNA binding sites, is one of the best-studied aspects of regulation. Gene regulatory networks (GRNs), which tabulate the interactions between TFs and their target genes, provide a succinct yet powerful framework to describe and model the regulation of a biological system (Davidson & Erwin, 2006; Marbach et al., 2016; Sinha et al., 2020). GRNs capture the architecture of gene regulation, revealing how regulators coordinate the expression of downstream target genes and how regulatory circuits can be dynamically reconfigured to drive cellular transitions (Aibar et al., 2017). Despite their importance, accurate inference of GRNs from high-throughput data remains

a fundamental problem in bioinformatics. First, regulatory relationships are dynamic and context-specific, meaning that the same TF may regulate different sets of target genes in different cell types and environmental or experimental conditions. This context-dependence makes it difficult to construct a universal GRN model and requires approaches that can capture the conditional regulatory interactions. Second, the experimental data used for GRN inference is characterized by high dimensionality, inherent noise, and technical artifacts that can obscure regulatory signals. Traditional bulk RNA sequencing (bulk RNA-seq) provides population-averaged measurements that may mask cell-to-cell differences (Lee & Tzou, 2009). Even with the advent of single-cell technologies, the data are still challenging to analyze due to high dropout rates and other forms of technical noise (Kharchenko et al., 2014). Third, discerning causal from correlative relationships poses additional challenges for GRN inference. Most computational approaches rely on statistical associations between the expression levels of TFs and their putative targets (Aibar et al., 2017; Johnson et al., 2007; Margolin et al., 2006), yet correlation does not imply causation. Disentangling direct regulatory interactions from indirect effects requires analytical approaches capable of modeling complex regulatory networks and accounting for potential confounding factors, potentially via use of multi-omics data (Bravo Gonzalez-Blas et al., 2023; Kim et al., 2023; Stuart et al., 2019). Lastly, when using multi-omics data to improve specificity, the heterogeneity, scale and complexity of data presents additional challenges. The integration of multiple omics modalities such as scRNA-seq and scATAC-seq enhances the comprehensiveness of GRN reconstruction, but it also introduces additional technical difficulties related to batch correction and joint modeling of different data types with distinct statistical properties (Bravo Gonzalez-Blas et al., 2023; Dibaieinia & Sinha, 2020; Stuart et al., 2019).

1.2 SPRED: A Simulator-Supervised Framework for GRN Reconstruction

This thesis presents a series of computational approaches for GRN inference that address different aspects of the above-mentioned challenges. In the first project (Chapter 2), we develop a novel simulator-supervised neural network framework, SPREd (Supervised Predictor of Regulatory Edges) (Wu & Sinha, 2024), for GRN reconstruction from transcriptomic data. This work addresses the task of GRN reconstruction from gene expression data alone, a common requirement for studies where the biological process under study has not yet been probed with epigenomics and other “omics” data apart from the relatively inexpensive transcriptomics modality. The dominant paradigm for expression-based GRN inference involves training a separate multi-variable model for each gene that can predict its expression using TFs as input features. The TFs found to be most important for the prediction are then assigned as the gene's regulators. One of the major problems is that the number of potential regulators (TFs) is often much larger than the number of samples, making it difficult to train models robustly. Additionally, it forces a pre-determined mathematical form (that encapsulated in the predictive model) onto the complex regulatory relationships. To overcome these limitations, recent methods have shifted toward training models that directly predict TF–target gene interactions. That is, the modeling task is changed from that of predicting a gene’s expression from TFs’ expression levels to that of predicting a TF-gene relationship from the expression levels of TFs as well as the target genes. However, this emerging alternative paradigm presents its own challenges due to the absence of large-scale, reliable ground truth as the labels of training data. These critical bottlenecks in GRN reconstruction form the motivation for our work. The SPREd model leverages synthetic data generation through a biophysics-based simulator, allowing for training deep neural networks that directly predict the relationships between TFs and target genes from expression data. Our simulator-supervised learning strategy exploits large-scale synthetic datasets that are easily created while capturing realistic regulatory dynamics. We test SPREd on diverse synthetic and real data sets,

demonstrating its improved accuracy in identifying both direct and indirect regulatory relationships compared to other GRN inference models.

1.3 Dissecting GRNs in Atherosclerosis with Single-Cell Multi-Omics

While SPREd enables us to learn the underlying regulatory logic from bulk transcriptomics data, the second part of this thesis focuses on the reconstruction of GRNs from single-cell multi-omics data in disease contexts. The ability to measure multiple types of molecular species and events, including mRNA concentrations, DNA accessibility and specific histone modifications, on genome-wide scale, offers an opportunity to dissect the regulatory landscape with remarkable resolution. This is particularly crucial for understanding heterogeneous biological systems consisting of changing cell types in different conditions, especially those involved in disease progression (Duan et al., 2023).

The research presented in Chapter 3 is motivated by the overarching goal of identifying the key TFs driving atherosclerosis, a chronic inflammatory disease of the arteries that can lead to heart attacks and strokes (Libby et al., 2009). The development of atherosclerosis involves complex changes in various cell types, including endothelial cells (ECs), smooth muscle cells and various immune cells. By integrating scRNA-seq and scATAC-seq data from multiple time points and experimental conditions relevant to atherosclerosis progression, we can gain a much more comprehensive view of the regulatory changes that cause atherogenesis. We use a combination of cis-regulatory analysis and coexpression-based GRN inference to identify TFs that regulate atherogenesis-linked transcriptomic shifts in ECs. We also develop a new method that incorporates cis-regulatory evidence as prior information in a neural network model of gene expression, which subsequently yields GRN edges through explainable machine learning techniques. We present a novel strategy for visualization of regulatory profiles of genes, by using supervised learning to map each gene to a low dimensional embedding,

allowing a global view of all genes that captures their differential expression patterns and regulatory evidence. Our systematic analysis provides insights into the regulatory programs controlling flow-induced reprogramming of ECs during the development of atherosclerosis. We predict the TFs Creb3l2, RelA, and Mef2c to coordinate the transition from proatherogenic ECs to pathological states, and validate these predictions in vitro through collaboration.

1.4 Summary

In conclusion, this thesis presents a suite of innovative computational strategies that addresses fundamental challenges in GRN inference. By designing a simulator-supervised learning framework and developing novel algorithms for the integrative analysis of single-cell multi-omics data, we have developed high-performance and flexible methods for dissecting the intricate regulatory programs that underpin cellular function and disease. The application of these methods to the study of atherosclerosis provides an example of how the synergy between principled computational innovation and deep biological investigation can lead to a deeper understanding of disease processes and facilitate the development of new therapeutic strategies.

CHAPTER 2. SPRED: A SIMULATION-SUPERVISED NEURAL NETWORK TOOL FOR GRN RECONSTRUCTION

This chapter is heavily excerpted from our paper published on *Bioinformatics Advances*. “Wu, Zijun, and Saurabh Sinha. "SPREd: a simulation-supervised neural network tool for gene regulatory network reconstruction." *Bioinformatics Advances* 4.1 (2024): vbae011.”

2.1 Abstract

Reconstruction of gene regulatory networks (GRNs) from expression data is a significant open problem. Common approaches train a machine learning (ML) model to predict a gene’s expression using transcription factors’ (TFs’) expression as features and designate important features/TFs as regulators of the gene. Here, we present an entirely different paradigm, where GRN edges are directly predicted by the ML model. The new approach, named “SPREd” is a simulation-supervised neural network for GRN inference. Its inputs comprise expression relationships (e.g., correlation, mutual information) between the target gene and each TF and between pairs of TFs. The output includes binary labels indicating whether each TF regulates the target gene. We train the neural network model using synthetic expression data generated by a biophysics-inspired simulation model that incorporates linear as well as non-linear TF-gene relationships and diverse GRN configurations. We show SPREd to outperform state-of-the-art GRN reconstruction tools GENIE3, ENNET, PORTIA and TIGRESS on synthetic datasets with high co-expression among TFs, similar to that seen in real data. A key advantage of the new approach is its robustness to relatively small numbers of conditions (columns) in the expression matrix,

which is a common problem faced by existing methods. Finally, we evaluate SPREd on real data sets in yeast that represent gold standard benchmarks of GRN reconstruction and show it to perform significantly better than or comparably to existing methods. In addition to its high accuracy and speed, SPREd marks a first step towards incorporating biophysics principles of gene regulation into ML-based approaches to GRN reconstruction.

2.2 Introduction

2.2.1 Understanding Gene Regulation Network

Gene regulatory networks (GRNs) are a popular framework for describing mechanisms underlying transcriptomic changes associated with a variety of biological processes, such as development (Peter & Davidson, 2011; Spitz & Furlong, 2012), behavior (Sinha et al., 2020) and cancer (Emad & Sinha, 2021). GRNs catalog transcription factors (TFs) and their target genes, each such TF-gene relationship represented as an edge in a network. Typically, such an edge means that the TF directly regulates the gene, i.e., perturbing the TF's concentration should change the gene's expression and that such a causal relationship arises from the TF binding to an enhancer or promoter associated with the gene. Accordingly, methods for reconstructing GRNs rely on statistical relationships between TF and gene expression, as well as evidence of TF-DNA binding that might underlie those relationships (Aibar et al., 2017; Johnson et al., 2007). However, it is often the case that the researcher only has access to gene expression data for their system, with other types of data such as TF-DNA binding profiles, histone modification profiles, and 3D chromatin data being unavailable (Traniello et al., 2023). Thus, GRN reconstruction solely from expression data is an important open problem, and the focus of our study.

Given a matrix of expression values, with rows representing genes (including TF genes) and columns representing different conditions or cells that have been profiled with

transcriptomics technologies (Wang et al., 2009), the task is to infer the underlying GRN. This amounts to detecting covariation between expression levels of TF and gene (rows) across the different measurements (columns), and the biological interpretation of an edge is thus contingent on the kind of variation represented by columns of the matrix. The nature of variation and covariation present in expression matrices differs substantially between single-cell and “bulk” transcriptomic datasets, leading to GRN inference methods specialized for either domain. For instance, methods for the single-cell domain must tackle high levels of technical noise (Kharchenko et al., 2014) but have very large numbers of samples (~1000 to ~100,000 cells) to rely on, while GRN inference from bulk transcriptomic data must work with few samples (~10 to ~100 conditions or individuals) but do not face the challenge of dropouts (Lee & Tzou, 2009). Here, we focus on GRN inference for bulk transcriptomics data, though the core ideas are germane to the single-cell domain as well.

2.2.2 Existing Approaches and Challenges in GRN Reconstruction

Many statistical and machine learning approaches have been proposed for GRN reconstruction from expression data, including those based on correlations and information theoretic measures (Chan et al., 2017; Stuart et al., 2003; Wagner et al., 2016), probabilistic graphical models including Bayesian networks (Friedman et al., 2000; Margolin et al., 2006; Markowitz & Spang, 2007; Perrin et al., 2003; Yu et al., 2004), Boolean networks (Liang & Han, 2012; Shmulevich et al., 2002), differential equations (Matsumoto et al., 2017), linear regression (Guo et al., 2016; Haury et al., 2012; Yip et al., 2010), Random Forests (Huynh-Thu & Geurts, 2018; Huynh-Thu et al., 2010), gradient boosting (Aibar et al., 2017; Lim et al., 2013; Slawek & Arodz, 2013), neural networks (Chen et al., 2021; Li et al., 2022; Lin & Ou-Yang, 2023; Wang et al., 2020; Yuan & Bar-Joseph, 2019; Zheng et al., 2022) among others. The more popular approaches are based on the idea of training a multi-variable model that predicts a gene’s expression from the levels of TFs, reporting

the TFs that are most useful for such prediction as the regulators of the gene (A separate model is trained for each gene). A major challenge faced by these methods is that the number of covariates is typically far greater than the number of samples, making it difficult to train the models robustly (Ruyssinck et al., 2014). A second, related limitation of the approach is that the mathematical “form” of the relationship between TFs’ and gene expression must be pre-determined and encoded into the structure of the model, e.g., as a weighted sum, decision tree, etc., and rich representations that accommodate a greater variety of possible relationships, e.g., neural networks, must be eschewed due to small sample size.

Recent studies have explored a different line of attack on the problem: that of training a model to directly predict a TF-gene relationship from expression profiles of a gene and its candidate regulator. The key difference here is that a “sample” (unit for which a prediction is to be made) is now a TF-gene pair rather than a biological condition, and the output of the model is the presence or absence of regulatory relationship between that TF-gene pair, rather than gene expression in a condition. This addresses the small sample size of above-mentioned methods, since the number of TF-gene pairs is large, but poses a different challenge: that of determining a training set where many TF-gene pairs are labeled as having true regulatory relationships or not. One possible solution, adopted by Yuan & Bar-Joseph (Yuan & Bar-Joseph, 2019), is to use epigenomic evidence of TF-gene relationships (e.g., TF ChIP-seq) to define positive examples for training. With a large enough training set of TF-gene pairs and their positive/negative labels, various classes of ML models, including deep neural networks (Lin & Ou-Yang, 2023; Yuan & Bar-Joseph), may then be trained to predict GRN edges from the joint distribution of TF and gene expression. A distinct advantage of this approach is that it offers greater flexibility in modeling the relationship between TF and target expression. In this study, we sought to further explore this emerging approach to GRN inference, which departs significantly from

the dominant paradigm today. We will refer to this approach as “supervised GRN reconstruction” since it relies on examples of true and false GRN edges to train predictive models.

Despite the promise of its early versions (Yuan & Bar-Joseph), a major limitation of the supervised GRN reconstruction strategy is in setting up the training set. This is because “gold standard” (highly accurate and comprehensive) GRNs are almost non-existent today. Indeed, the most widely adopted strategy for benchmarking GRN inference tools, including in community-wide efforts such as DREAM challenges (Marbach et al., 2012), continues to rely on synthetic data sets. “Real data” benchmarks are limited to two or three GRNs on which all inference tools exhibit very low accuracy levels, due in no small part to the incompleteness of those GRNs. On the other hand, there has been a surge in development of realistic simulators of expression data, including those that simulate the dynamics of a GRN using biophysical principles and incorporate realistic models of expression noise (Dibacina & Sinha; Wagner et al., 2016). This presents the opportunity of utilizing simulated data sets and their underlying GRNs as training data for supervised GRN reconstruction. This is the key idea explored in our work.

2.2.3 SPREd: A Supervised Solution Leveraging Simulated Data

Motivated by the above considerations, we develop here a new GRN inference tool called SPREd, which trains a neural network model to predict GRN edges from expression profiles. It relies on a very large training set comprising millions of TF-gene pairs, generated using a state-of-the-art expression simulator. Systematic evaluations show that SPREd outperforms leading methods such as GENIE3, ENNET, TIGRESS and PORTIA on unseen synthetic data sets. We also find that SPREd generalizes well when tested on synthetic data with different characteristics from those used in training. Finally, we

demonstrate that SPREd is more accurate than other methods on the most reliable and widely used “real data” benchmark available today. We believe SPREd is a promising first step towards more advanced implementations of supervised GRN reconstruction where our rapidly developing understanding of GRN dynamics are encoded in increasingly realistic simulators, which then lead to even more powerful training sets for training highly expressive ML models.

2.3 Methods

2.3.1 *Simulation of Expression Matrix from a GRN, using SERGIO*

The SERGIO simulator (Dibaenia & Sinha, 2020) requires “targets file” describing the regulators of each gene in the GRN and the “master regulators file” with information on MRs as inputs. The targets file lists the non-MR gene’s identifier, the number of regulators of that gene, their identifiers, interaction parameter (K) representing the strength and directionality of influence of corresponding regulators, and Hill coefficients representing the degree of non-linearity of the influence (see SERGIO documentation). Each K was sampled from a uniform distribution on $[1.0, 5.0]$, and then negated with 0.2 probability. Each Hill coefficient was set to 2 with 0.9 probability and 1 with 0.1 probability. The number of “cell types” to be simulated was set to 100, and the number of cells per cell type was set to 1. The master regulators file requires specification of each MR’s production rate in each cell type. This was done by first selecting, for each GRN, a pair of ranges defining low and high rates respectively, and then sampling each MR’s rate in each cell type at uniform from these ranges. (The pair of ranges were as suggested by SERGIO and chosen uniformly at random.) Given the targets file and master regulators

file, the SERGIO simulator produces steady-state expression matrices with dimension of $(n_{MR} + n_{TF} + n_G) \times 100$. Each of the m “cell types” simulated by SERGIO was treated as a separate “condition”. In benchmarks where $m < 100$, m of the 100 simulated “cells” were selected at random as conditions.

2.3.2 Performance Evaluation Metrics

All evaluations were performed for each target gene separately and averaged over genes. For a gene, the task is to predict whether each of the n_{TF} candidate regulators is actually a regulator, i.e., a binary classification task on n_{TF} test samples, allowing computation of AUROC. Instead of AUPRC, we used the related metric “Average precision” (Xu et al., 2020) $AP = \sum_n (R_n - R_{n-1})P_n$, where R_n and P_n denote the n^{th} threshold of recall and precision. GRNs in benchmarks used here have very few regulators per gene, which may lead to misleading AUPRC values, especially in the extreme but common case of one true regulator.

2.3.3 Input Data Preparation in SPREd

Given an expression matrix with n_{TF} rows representing TFs and one row for the target gene, and m columns representing conditions, the first step is to use a Box-Cox power transformation (Bagnara, 1964) on each gene, followed by z-transformation resulting in zero mean and unit standard deviation for each gene. (In case the expression matrix has any negative values, all entries are shifted by a constant to achieve strictly positive values.) Pairs of rows of the resulting normalized expression matrix were examined to calculate a covariance matrix, denoted as Σ , Pearson correlation coefficient matrix, Spearman correlation coefficient matrix, and discrete mutual information matrix, each of

dimensionality $N \times N$, where $N = n_{TF} + 1$. The $N \times N$ precision matrix was calculated using $(\Sigma + \varepsilon \times I)^{-1}$, $\varepsilon = 1e - 3$. Since the resulting matrices are symmetric, $\binom{N}{2} + N$ distinct entries of each matrix are used as features for the SPREd-ML model. The SPREd-SP model, on the other hand, predicts only whether the i^{th} TF is a regulator of the target gene or not, and its features include row i of the above five matrices. For both models, all pairwise relationship features of each of the five types are flattened into a vector.

2.3.4 Architecture of SPREd-ML

The SPREd-ML architecture is shown in Supplementary Figure 8. The input of the neural network is a $n_{pairs} \times 5$ matrix as described above. The n_{pairs} rows of this matrix include $\binom{N}{2}$ rows representing each TF-TF pair and N rows representing each TF-target gene pair, the target gene being fixed. Each of the former set of $\binom{N}{2}$ rows feeds into a shared 1D convolutional kernel, while each of the latter set of N rows feeds into another shared 1D convolutional kernel. These two kernels result in feature maps of dimensions $\binom{N}{2} \times 1$ and $N \times 1$ respectively, which are fully connected to n_{TF} units of the output layer. We used rectified linear activation (ReLU) as the activation function throughout the network. Connections between the feature map representing TF-target gene pairs and the output layer had a dropout rate of 0.3.

2.3.5 Architecture of SPREd-SP

The overall architecture is presented in Figure 2.1B. The input is a $n_{pairs} \times 5$ matrix, with the $n_{pairs} = N = n_{TF} + 1$ rows including n_{TF} rows for TF-TF pairs (one of the TFs being fixed) and one row for the TF-target gene pair. These inputs feed into two

1D convolution-batchnorm-ReLU blocks with dimensions 5×16 and 16×1 respectively. (The n_{TF} rows for TF-TF pairs and the single row for the TF-gene pair are convolved using separate kernels.) The resulting feature map is a n_{pairs} -dimensional vector that is fully connected to a 128-node layer, which in turn is fully connected with the single output node.

2.3.6 *SPREd Training Details*

Both models are trained using Adam optimizer with weight decay of $5e-4$ and learning rate of $2e-4$. Binary cross entropy with logits is used as the loss function. The positive weight is set to 9, the batch size is set to 32 and the maximum number of training epochs is set to 300. The training was performed on an Nvidia V100 GPU. See Supplementary Table 4 for samples of run-times of training and application of trained models (testing) of SPREd.

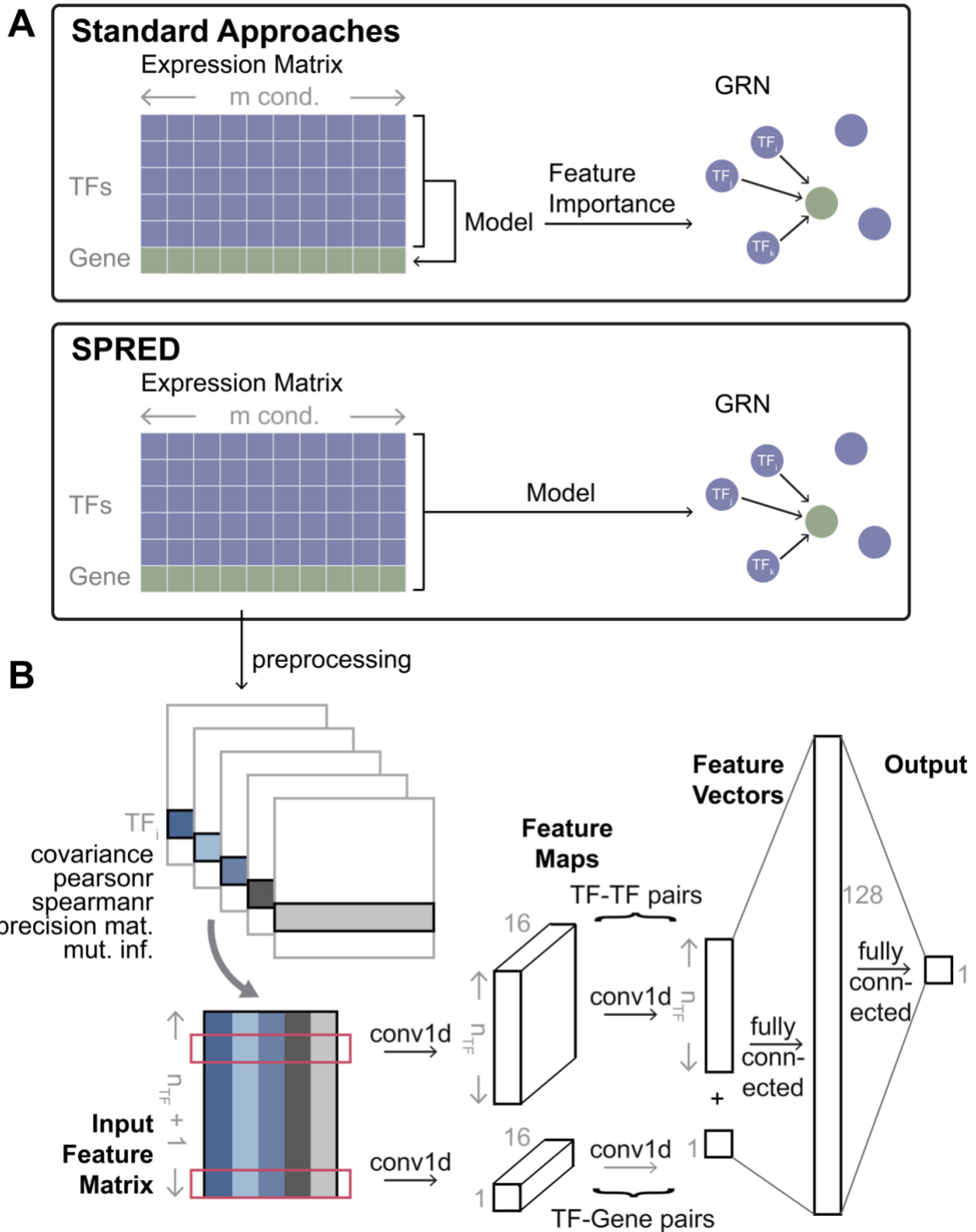


Figure 2.1 SPRED: a simulation-supervised learning framework for gene regulatory network (GRN) inference. (A) Standard approaches typically build ML models of the target genes using the expression levels of TFs as features. GRNs are then constructed based on the feature importance of TFs (features) in the model trained for a target gene. In SPRED, an ML model is trained to directly predict TFs regulating a target

gene, based on expression matrix of all TFs and the target gene. The ML model is trained on simulated expression matrix-GRN pairs and can then be used to predict the GRN for any expression matrix. (B) Architecture of SPRED-SP neural network model. Given an expression matrix whose rows represent n_{TF} TFs and one target gene (panel A), the preprocessing step creates five matrices of pairwise relations (features) for each TF-TF pair and each TF-target gene pair. These features include covariance, Pearson correlation, Spearman correlation, mutual information and precision matrix entry corresponding to the TF-TF or TF-target gene pair. The five features of every gene pair involving a particular TF (say TF_i) then serve as the inputs of a 1D convolutional neural network (CNN) (input feature matrix, size $(n_{TF}+1) \times 5$). The feature map resulting from the first layer of convolution (out channel = 16) is of dimension $(n_{TF}+1) \times 16$, and feeds into a second convolution layer, whose outputs are fully connected to a hidden layer, which finally connect to the output layer. The output layer consists of a binary label indicating if TF_i is a regulator of the target gene (details shown in Methods).

2.4 Results

2.4.1 SPRED: Supervised Learning Framework For GRN

The GRN inference problem addressed here is to detect the TFs regulating a given gene based on expression data across a range of “conditions”. (We will use “conditions” to refer to different experimental conditions, replicates, cell types or cells.) Leading methods use ML for this task, modeling the expression of the target gene as a function of the expression levels of all TFs, then examining the contribution of each TF (feature) to the model and reporting the important TFs as regulators of the gene (Figure 2.1A top). In other words, current methods approach GRN prediction via a related problem – gene expression prediction – and extract the GRN as a “by-product” of solving that problem. In contrast, we use a more direct, supervised learning approach to GRN prediction (Figure 2.1A bottom). Instead of relating TFs’ expression (input) to target gene expression (output), we relate the TFs’ and target gene’s expression (input) to TF-gene edges (output). Learning

such a relationship requires training data comprising many input-output pairs, each of which comprises an expression matrix (input) and its corresponding GRN (output). Since reliable GRNs are rarely available, we used a simulator that constructs a realistic but synthetic expression matrix corresponding to a GRN, repeating the process many times, and used the resulting expression matrix-GRN pairs to train the GRN prediction model. Once such a model has been trained, it can be presented expression data that has not been seen before and directly predict the GRN that those data may have arisen from.

Our method, called SPREd (Supervised Predictor of Regulatory Edges), utilizes a neural network to relate an expression matrix to the corresponding GRN. Here, we will assume that the expression matrix includes the expression levels of a single target gene and all the candidate TFs across a set of m conditions, and the GRN consists of a subset of the n_{TF} candidate TFs that are regulators of the target gene (Figure 2.1A). We implemented two complementary models, “SPREd-MultiLabel” (SPREd-ML) and “SPREd-SinglePair” (SPREd-SP), for the GRN prediction task. These two models have similar architectures and feature definitions but differ in one key aspect: SPREd-SP predicts regulatory relationships one TF-gene pair at a time, while SPREd-ML simultaneously scores all n_{TF} candidate TFs for their regulatory relationship to a gene. For both models, the input expression matrix is represented via features that capture the relationships between TF(s)-target gene pairs and TF-TF pairs. For SPREd-SP, when predicting the relationship between TF_i and gene g , features are calculated for the pair (i, g) and (i, j) for every candidate TF_j , yielding a total of $n_{pairs} = 1 + n_{TF}$. For SPREd-ML, since the output of the model includes relationships between every candidate TF and gene g , features are calculated for all n_{TF} TFs paired with gene g and for all $\binom{n_{TF}}{2}$ TF-TF pairs, yielding

$n_{pairs} = n_{TF} + \binom{n_{TF}}{2}$. For each pair, five statistics including covariance, Pearson correlation, Spearman correlation, mutual information, and “precision score” are calculated (see Methods). These five statistics for each of the TF-gene and TF-TF pairs are presented as a $n_{pairs} \times 5$ matrix of “features” in the input layer of a 1d convolutional neural network (CNN). The architecture of the neural network (Figure 2.1B and Supplementary Figure 8), with convolution layer(s) producing a n_{pairs} -dimensional vector (convolving the five features of each pair into one value), followed by fully connected hidden layers to the output layer, which consists of one unit for SPREd-SP or n_{TF} units for SPREd-ML with values of 0-1 indicating the absence or presence of an edge between the TF and the target gene.

The SPREd neural network is trained on synthetic data generated using the SERGIO simulator. SERGIO (Figure 2.2A) takes an input GRN, with a subset of TFs marked as master regulators (MRs) that are TFs without input regulators thereof. Starting with the expression levels of the MRs, SERGIO simulates the dynamics of transcriptional regulation and samples cells from the steady state distributions of those dynamics. By varying the MRs’ expression profiles, we can diversify the biological conditions simulated by SERGIO. To generate training data for the SPREd model, we created GRNs with a special structure (Figure 2.2A, left). Each GRN has a set of n_G target genes, regulated by a set of n_{TF} TFs, which are in turn regulated by n_{MR} MRs. A gene has $d_{TF \rightarrow g}$ incoming regulatory edges, while each TF is under the control of $d_{MR \rightarrow TF}$ MRs. This simple 3-layer structure allows us to control the complexity of combinatorial regulation and global co-expression in the simulated expression data used for model training and testing. To generate an expression matrix of m conditions, we used SERGIO to simulate m cells

specified by m distinct MR profile (see Methods). This produces an expression matrix with $n_{MR} + n_{TF} + n_G$ rows and m columns, which is then processed to generate training samples for SPREd (see Methods), with $n_{TF}n_G$ samples being created for SPREd-SP and n_G samples for SPREd-ML. The whole process is repeated n_{GRN} times, each time creating a GRN at random with the desired connectivity properties.

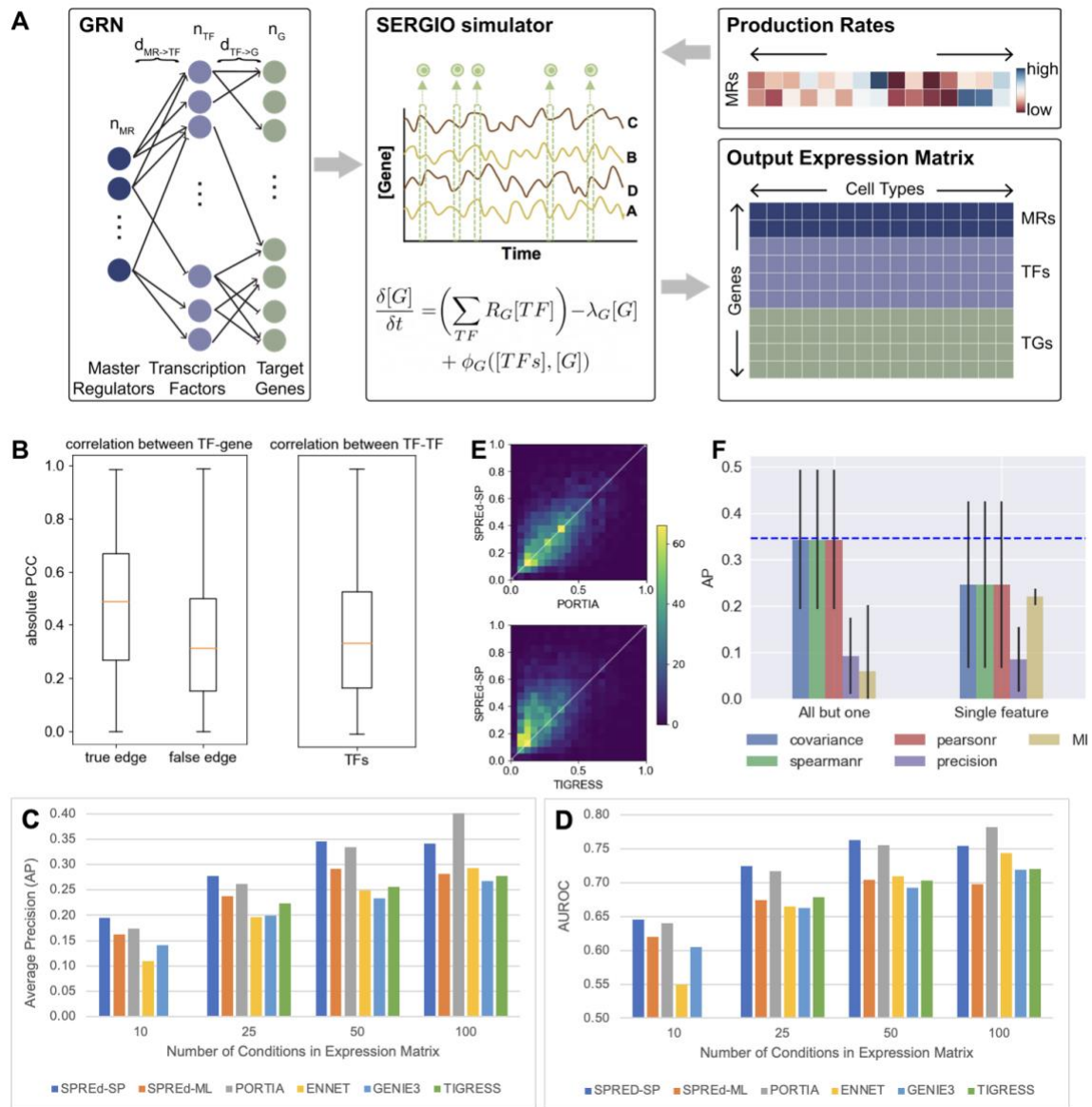


Figure 2.2 SPREd exhibits superior performance on synthetic datasets. (A) Schematic of synthetic data generation. Each synthetic data set comprises a GRN

(left) and an expression matrix (bottom right). The GRN has three “layers” – master regulators (MR), transcription factors (TF) and target genes, with regulatory edges from one layer to the next. The MRs are included as the first layer so as to induce co-expression among TFs, mimicking real data. Parameters describing the GRN include the number of MRs (n_{MR}), the number of transcription factors (n_{TF}), the number of target genes (n_G), the number of incoming edges to each target gene ($d_{TF \rightarrow G}$), and the number of incoming edges to each TF ($d_{MR \rightarrow TF}$). A GRN is sampled at random while respecting these parameters and is used by SERGIO (middle), a biophysics-based model, to simulate the expression profiles of different artificial biological conditions, each of which is described by the production rates of MRs (top right), thus generating an expression matrix whose rows include target genes, TFs and MRs and columns represent biological conditions. (B) Co-expression statistics of synthetic expression data. Absolute value of Pearson correlation coefficient (PCC) of TF-gene pairs (left) that comprise GRN edges (“true edges”) and those that do not (“false edges”) and of TF-TF pairs (right). Results are from simulations using GRNs in default configuration. (C,D) Average precision or “AP” (C) and AUROC (D) of the six evaluated methods – SPREd-SP, SPREd-ML, PORTIA, ENNET, GENIE3 and TIGRESS – on data sets with varying numbers of conditions (columns in expression matrix). Each performance metric (AP or AUROC) is calculated for individual target genes, and results summarized over 5000 genes from 50 GRNs (100 genes in each GRN). (E) Direct comparison of average precision (AP) between SPREd and PORTIA (top) or TIGRESS (bottom), for expression data with 50 conditions. (F) AP of SPREd-SP when using all but one (left) or only one (right) of the five features describing each TF-gene or TF-TF pair. AP when using all five features is shown by blue dashed line.

2.4.2 *SPREd Outperforms Leading GRN Inference Methods on Synthetic Benchmarks*

We first evaluated the prediction accuracy of SPREd on a synthetic benchmark that resembles the training data described above. This benchmark comprises 50 GRNs, each with $n_G = 100$ target genes, and thus presents $50 \times 100 = 5000$ “test samples”, each being a GRN with a single target gene and its known regulator TFs, and the corresponding expression matrix to be used as input. A GRN reconstruction method must predict, based on the expression matrix for each test sample, which of the n_{TF} candidate regulators actually regulate the target gene. In general, such methods assign scores to each candidate regulator and their accuracy can be assessed using standard metrics such as Area Under

Receiver Operating Characteristic (AUROC) or Area Under Precision Recall Curve (AUPRC). Here, in addition to the AUROC, we relied on the Average Precision (AP, see Methods) score rather than the AUPRC, because it is better suited for extremely sparse positive sets. As comparators for evaluating predictive accuracy, we relied on four state-of-the-art methods – ENNET (Slawek & Arodz, 2013), GENIE3 (Huynh-Thu et al., 2010), TIGRESS (Haury et al., 2012) and PORTIA (Passemiers et al., 2022), which have emerged as top methods in recent benchmarking studies (Passemiers et al.) and represent a diversity of modeling approaches.

In the first benchmark, we set the number of conditions m to 50 and the GRN parameters to $n_{MR} = 5$, $d_{MR \rightarrow TF} = 3$, $n_{TF} = 100$ and $d_{TF \rightarrow G}$ in the range [3,7]. The number of conditions is typical for bulk transcriptomics data from an individual PI’s laboratory (a few experimental conditions and 3-5 replicates per condition (Saul et al.)). It is also the ballpark number of samples (individuals) for each tissue in GTEx (Consortium, 2013). The parameters n_{MR} and $d_{MR \rightarrow TF}$ determine the level of co-expression among TFs, an important characteristic of the benchmark since extensive co-expression among TFs in real data sets is one of the major hurdles in accurate GRN prediction. Figure 2.2B shows the extent of TF-TF correlation in the generated benchmark, which is comparable to that of TF-gene pairs that are not GRN edges; GRN edges have significantly greater correlations on average, as expected, but we also note a substantial overlap between the distributions of edges and non-edges, pointing to the difficulty of the GRN prediction task for this benchmark. We trained SPREd on a data set of $n_{GRN} = 250$ GRNs generated with the same simulation parameters as the test set. As shown in Figure 2.2C (group “50”) and Table 2.1 A ($m = 50$), the AP of SPREd-SP is higher on average than those of all other

methods, with strong statistical significance in most cases (Supplementary Table 1A). Similarly, SPREd-SP has a higher AUROC than these methods except for PORTIA, to which it is comparable (Figure 2.2D, Table 2.1B, Supplementary Table 1B). Figure 2.2E shows head-to-head comparisons of SPREd-SP with the two next-best competing methods (PORTIA and TIGRESS), revealing higher AP score of SPREd-SP versus TIGRESS for the majority of test samples, and a noticeable degree of complementarity between SPREd-SP and PORTIA. We noted SPREd-ML to have weaker performance than SPREd-SP and PORTIA in these evaluations, though it exhibited higher AP and AUROC than ENNET, GENIE3 and TIGRESS.

Table 2.1 Performance comparison of SPREd, PORTIA, ENNET, GENIE3 and TIGRESS. Average Precision (AP) (A) and AUROC (B) in evaluations on expression matrices with varying number of conditions (columns in expression matrix). Each performance metric (AP or AUROC) is calculated for individual target genes, and results averaged over 5000 genes from 50 GRNs (100 genes in each GRN). Expression data are simulated using GRNs with $n_{MR}=5$, $n_{TF}=100$, $d_{TF \rightarrow G} = 3-7$. Note: TIGRESS runs for 10 conditions did not complete successfully.

(A) AP

#conditions	SPREd-SP	SPREd-ML	PORTIA	ENNET	GENIE3	TIGRESS
10	0.19	0.16	0.17	0.11	0.14	-
25	0.28	0.24	0.26	0.20	0.20	0.22
50	0.35	0.29	0.33	0.25	0.23	0.26
100	0.34	0.28	0.40	0.29	0.27	0.28

(B) AUROC

#conditions	SPREd-SP	SPREd-ML	PORTIA	ENNET	GENIE3	TIGRESS
10	0.65	0.62	0.64	0.55	0.61	-
25	0.72	0.67	0.72	0.67	0.66	0.68
50	0.76	0.70	0.76	0.71	0.69	0.70
100	0.75	0.70	0.78	0.74	0.72	0.72

We next repeated the above evaluations at additional values of m (number of conditions) and observed the performance advantage of SPREd-SP to persist when there are fewer conditions ($m = 25, 10$) (Table 2.1, Figure 2.2C,D, Supplementary Table 1), though with $m = 100$, the performance of PORTIA improves over SPREd-SP. Similarly, the above-noted advantage of SPREd-ML over ENNET, GENIE3 and TIGRESS persists on data with fewer conditions but disappears at $m = 100$. (Also see Supplementary Figure 1). This is in line with our expectation that with sufficient sample size to train models of gene expression, the need for an alternative paradigm such as the synthetic data-supervised approach of SPREd diminishes. In this regime, the ability to tune free parameters on the given expression matrix can give an edge to the standard approach (Figure 2.1A) in vogue today. We note that where this regime begins (e.g., ~ 100 conditions in our benchmark) depends on the data set and has to be estimated with care.

We also observed similar relative performances when comparing SPREd-SP with other tools on the DREAM5 benchmark of (Marbach et al., 2012), see Supplementary Table 2 and Supplementary Table 3. In particular, we observed SPREd performance to be the best among compared methods in terms of AP, for three of the four networks, the exception being the network with the largest ratio of number of conditions (samples) to TFs (covariates).

We next determined the importance of the five features representing each gene pair to the accuracy of SPREd-SP. Firstly, we repeated the evaluations using only one of the features and found that the performance is substantially below that of the full model for any one of the features (Figure 2.2F, “Single feature”). This suggested that the model utilizes complementary information from multiple features. We also evaluated subsets

comprising four of the features, leaving out one at a time, and observed a substantial drop only when leaving out precision (Figure 2.2F, “All but one”), indicating that the information carried by each of the remaining features is largely redundant with at least one other feature (e.g., Pearson correlation and Spearman correlation are expected to be largely mutually redundant), whereas the precision score likely captures information complementary to the others. (See Supplementary Figure 2 for similar analysis with SPREd-ML.)

2.4.3 *Effect of Benchmark Diversity on Predictive Accuracy*

We next examined how SPREd performance varies with benchmark generation parameters that determine the difficulty of the GRN inference task. First, we looked at the effect of edge density between TFs and genes ($d_{TF \rightarrow G}$), set to 3 – 7 above. We repeated the evaluations with the range changed to [1,2] (less combinatorial regulation) or to [8,10] (more combinatorial). One might expect that as the edge density increases, the AP score should increase: if a gene has many regulators, then even a randomly predicted regulator is likely to be correct. However, we observed (Figure 2.3A) that the AP is highest in the $d_{TF \rightarrow G} \in [1,2]$ (least combinatorial) setting. (Also see Supplementary Figure 3 and Supplementary Figure 4.) One possible reason for this is that with greater numbers of true regulators of a gene, mutual correlations among TFs (Figure 2.2B) exacerbate the multicollinearity problem, making the TF prediction task harder. The literature suggests that the more complex cases of combinatorial transcriptional regulation (e.g., in early development) involve 3 – 7 TFs regulating a gene (Kazemian et al., 2010), but the rare examples of comprehensively reconstructed GRNs (e.g., in yeast) suggest an edge density

of 1 – 2 on average. These observations provide some real biological context to the evaluations above.

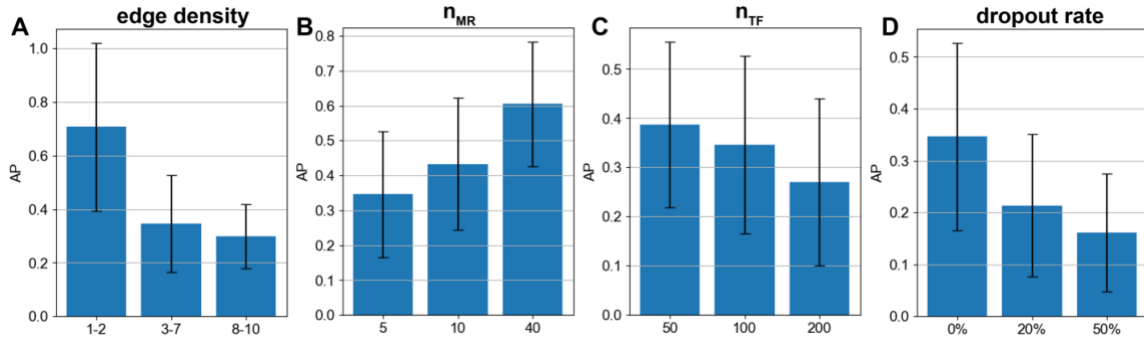


Figure 2.3. Effect of benchmark parameters on GRN reconstruction. Performance (average precision) of SPRED-SP with varying edge density ($d_{TF \rightarrow G}$) of 1-2, 3-7, and 8-10 TFs per target gene (A), varying numbers of MRs (n_{MR}) (B), varying numbers of TFs (n_{TF}) (C), and varying levels of dropout added to the synthetic expression matrix (D).

Recall that our motivation in adding the “master regulator” layer in synthetic GRNs is that it provides an easy way to control the extent of TF-TF correlations, a major challenge for GRN recovery. (In a more realistic, “cascade-like” GRN, two TFs’ mutual correlation will depend on their relative position in the cascade.) In our 3-layer GRNs, as n_{MR} increases, the pairwise correlation of TFs decreases and we found the GRN reconstruction to become easier, as expected (Figure 2.3B).

The parameter n_{TF} directly controls the difficulty of the GRN reconstruction task, since the task involves selecting the true regulators among all n_{TF} candidate TFs. Clearly, if we increase n_{TF} while keeping $d_{TF \rightarrow G}$ fixed, then for each gene the number of positives to be determined remains the same but the number of candidates increases, so the AP will decrease, which is what we observed (Figure 2.3C). We noted however less than the expected 2-fold decrease in AP when we double the number of TFs, from 50 to 100 and

then from 100 to 200. This suggests increasing n_{TF} leads to an unexplained improvement in ability to identify true regulators, countered by a stronger deterioration due to the larger pool of candidate TFs.

Finally, we examined the effect of expression measurement noise beyond the stochasticity of gene expression already modeled in the SERGIO simulation process. In particular, we simulated a simplistic version of “dropout” (Shrivastava et al., 2022) whereby a certain fraction of entries of the expression matrix are artificially at zero values. As expected, the performance deteriorates with increasing noise (Figure 2.3D) and at a dropout level of 50% the performance of SPRED-SP drops to nearly random expectation (Supplementary Figure S4). This corroborates previous reports that dropouts, unless properly mitigated by imputation tools (Eraslan et al., 2019; Li & Li, 2018; van Dijk et al., 2018), can be a major problem for GRN reconstruction (Dibaeinia & Sinha).

In the above evaluations, the two SPRED models, unlike the other four methods, are supervised with training data that has the same characteristics as the test data, which might give them an advantage. To partially offset this advantage, we repeated the comparisons on more heterogeneous benchmarks, with GRNs in the test set being created with a mixture of all settings of n_{MR} or $d_{TF \rightarrow G}$ parameters. We found (Figure 2.4B) that the advantages of SPRED models over comparators, seen above for mid-range edge densities (Table 2.1), are as strong or greater with the broader spectrum of edge densities. Similarly, when benchmarking with a broader span of n_{MR} values (5, 10, 40) that induce varying levels of TF-TF correlations, the advantage of SPRED over PORTIA and GENIE3 persists (Figure 2.4A), though the gap between SPRED and ENNET/TIGRESS shrinks. We repeated these evaluations with SPRED trained on GRNs with the benchmark parameter

(n_{MR} or $d_{TF \rightarrow G}$) set to a specific value while the test set remained a heterogeneous mix of GRNs, and found the model’s performance to remain unchanged (Supplementary Figure 5A,B). These results indicate that the performance of the model depends primarily on the difficulty of the test set rather than the precise characteristics of the GRNs used in training, which is a sign of generalizability.

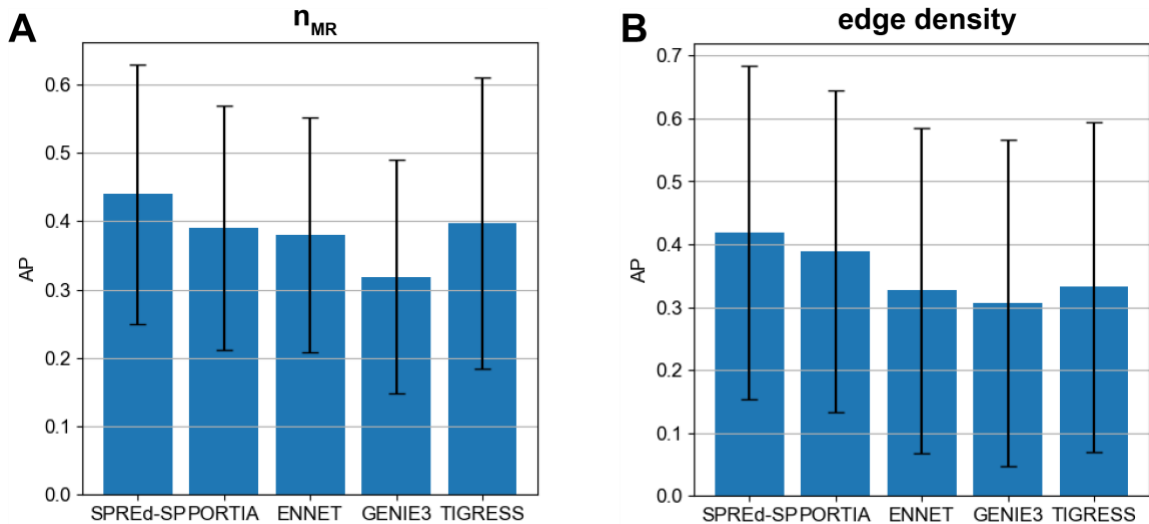


Figure 2.4. Performance comparison on heterogeneous benchmarks. AP comparison of SPRED-SP, SPRED-ML, PORTIA, ENNET, GENIE3 and TIGRESS on heterogeneous datasets comprising GRNs with (A) n_{MR} =5, 10 and 40 (in equal numbers) and (B) with $d_{TF \rightarrow G}$ = 1, 2, ... 10 (in equal numbers).

2.4.4 SPREd Performance is Better than or Competitive with Leading Methods on Real Data

Finally, we performed comparative evaluations of the six methods on one of the best characterized eukaryotic GRNs available today, from the yeast *S. cerevisiae*, reconstructed based on multiple types of cis-regulatory information, and compiled in the form of benchmarks by Siahpirani & Roy (Siahpirani & Roy). The Siahpirani-Roy yeast benchmark includes three “gold standard GRNs” named “MACISAAC2” (MacIsaac et al.,

2006), “YEASTRACT TYPE2” AND “YEASTRACT COUNT3” (see Methods), based on cis-regulatory evidence including TF ChIP-seq profiles, TF motif matches and evolution conservation and have since been used by other studies (Passemiers et al., 2022). The benchmark also includes three different expression matrices, named “NatVar”, “Knockout”, and “StressResp”, reflecting expression variation under different axes of biological differences. We evaluated the six methods for each of the nine combinations of the expression matrix as input and the gold standard GRN as ground truth, called “tests” below. Each test involves identifying typically 1 – 2 TFs regulating a target gene (out of ~250 candidate TFs), for ~2,000 target genes.

Table 2.2. Mean AP (A) and AUROC (B) of six evaluated methods on benchmarks using the MacIsaac2, YEASTRACT type2 (YEASTRACT2), or YEASTRACT count3 (YEASTRACT3) gold standard networks and the Nat Var (NV), Knockout (KO), and StressResp (SR) expression data from Siahpirani & Roy. AP/AUROC values are calculated for individual genes and averaged over genes. Highest value in each column is shown in bold. Median value across the nine benchmarks is shown for each method in the rightmost column.

(A) AP

	MacISAAC			YEASTRACT2			YEASTRACT3			Median
	NV	KO	SR	NV	KO	SR	NV	KO	SR	
SPREd-SP	0.05	0.03	0.03	0.06	0.03	0.04	0.07	0.06	0.06	0.050
SPREd-ML	0.05	0.03	0.06	0.05	0.06	0.04	0.06	0.05	0.03	0.047
PORTIA	0.07	0.03	0.04	0.08	0.03	0.04	0.08	0.04	0.04	0.039
TIGRESS	0.02	0.02	0.02	0.04	0.04	0.04	0.06	0.04	0.02	0.040
ENNET	0.04	0.03	0.03	0.05	0.04	0.04	0.05	0.04	0.04	0.038
GENIE3	0.04	0.03	0.03	0.06	0.03	0.04	0.06	0.03	0.04	0.038

(B) AUROC

	MacISAAC			YEASTRACT2			YEASTRACT3			Median
	NV	KO	SR	NV	KO	SR	NV	KO	SR	
SPREd-SP	0.54	0.53	0.48	0.56	0.54	0.52	0.55	0.55	0.53	0.544
SPREd-ML	0.56	0.55	0.60	0.59	0.62	0.59	0.57	0.59	0.52	0.587
PORTIA	0.55	0.52	0.51	0.59	0.52	0.52	0.58	0.52	0.52	0.523
TIGRESS	0.45	0.45	0.44	0.45	0.45	0.45	0.47	0.46	0.46	0.453
ENNET	0.53	0.52	0.52	0.54	0.54	0.52	0.53	0.54	0.52	0.530
GENIE3	0.55	0.53	0.51	0.57	0.51	0.52	0.58	0.52	0.51	0.525

First, we examined AP values (Table 2.2A) and noted that the SPREd models have the highest median performance across the nine tests, with either one yielding the highest AP on six of the tests and PORTIA exhibiting the greatest AP on the remaining three. A head-to-head comparison (Supplementary Figure 6) reveals that both models clearly perform better than or similar to ENNET, GENIE3 and TIGRESS across the tests, while both models are competitive with PORTIA with benchmark-dependent complementarity. Examination of AUROC values (Table 2.2B) revealed that the two SPREd models have the highest median performance across tests, with SPREd-ML being the best among all methods for seven of the tests and SPREd-SP and GENIE3 being the best in one test each. Supplementary Figure 7 reveals the clear advantage of SPREd-ML over all other methods, while SPREd-SP outperforms ENNET and TIGRESS and is competitive with PORTIA and GENIE3.

We noted that the AP values in Table 2.2A are relatively low (in the 5% range) in absolute terms, regardless of the method. This is consistent with previous studies (Passemiers et al., 2022) and is likely due to the sheer difficulty of the problem where only a handful of ~ 250 candidate regulators are known to be positives. Indeed, when we reexamined the AP values in light of the random baseline, we found up to 2.4-fold higher than expected AP values for SPREd predictions (Supplementary Table 6). For example, in the three tests noted where SPREd-ML achieves an AP of ~ 0.06 , this value is 2.1-fold, 2.3-fold and 2.4-fold greater than random. At the same time, the relatively low fold changes (in AP) over random expectation in several of the other tests suggests that even the best methods available today leave ample scope for improvement. It is also likely that the true regulators of a typical gene are largely missing from the gold standard GRNs and the precision values are underrepresented as a result.

2.5 Discussion

SPREd in its current form is designed and tested for GRN inference from bulk RNA-seq data, which, despite the spreading popularity of single cell technologies, remains the platform of choice when a number of samples in varying biological conditions have to be profiled, especially on average laboratory budgets but also in consortium efforts (Consortium). As such, our performance comparisons have been focused on state-of-the-art tools for the bulk data domain, which form an actively growing genre in its own right. Despite obvious parallels, GRN inference for bulk and single-cell transcriptomics data presents distinct challenges – far smaller sample sizes in the bulk case and far greater technical noise in the single-cell case. They also address two different biological notions of GRN: single-cell GRNs explain gene expression variation from cell to cell

(Kolodziejczyk et al., 2015), which may in part be due to cell type differences, while bulk-based GRNs explain expression variation across biological conditions or individuals. An obvious task cut out for the future is to explore the SPREd approach for single-cell GRN inference; such efforts may begin with training models on synthetic single-cell data generated using SERGIO in ‘noisy’ mode but will likely need to explore alternative model architectures as well.

Several avenues of future improvements to SPREd present themselves. The current model uses as features five different types of pairwise relationships between genes, three of which (covariance, Pearson correlation, Precision matrix) capture linear relationships, one (Spearman correlation) captures monotonic relationships and one (mutual information) is more general. It will be interesting to explore if a 2D convolution of the joint distribution between two genes, akin to the approach of (Yuan & Bar-Joseph), can improve the featurization of pairwise relationships used by SPREd’s neural network. Finally, given that SPREd does not perform parameter training, which is a strategy successfully used by its competing methods, a combination of these two complementary approaches will be a promising direction for future research.

GRN reconstruction from expression data is a very challenging problem, in part because the goal is one of causal TF-gene relationships based on observational data alone, and also because the structure of real GRNs leads to extensive correlations among the TFs themselves. To overcome these fundamental obstacles, other kinds of evidence such as cis-regulatory elements or TF knockout/knockdown data are used in influential studies (Bravo Gonzalez-Blas et al., 2023; Kamimoto et al., 2023; Siahpirani & Roy, 2017; Skok Gibbs et al., 2022), but such data may not always be easily available, which necessitates continued

attention to the core problem of expression-to-GRN mapping. A second major challenge facing GRN reconstruction is the scarcity of reliable “gold standard” GRNs corresponding to the available expression data. A GRN edge (TF-gene pair), to be considered true, should at the very least be supported by evidence of gene dysregulation upon perturbing the TF’s expression and by credible evidence of functional TF binding to an enhancer linked with that gene. To our knowledge, there are very few, if any, GRNs that meet these stringent criteria of credibility and are reasonably complete (i.e., many or most edges known). This is the reason why synthetic benchmarks remain the primary evaluation approach today (Dibaenia & Sinha, 2020; Li et al., 2023; Pratapa et al., 2020). At the same time, it is expected that method developers evaluate their GRN reconstruction tools on real data sets, but one should not be surprised if such evaluations reveal low accuracies, given the shortcomings of current GRN benchmarks in terms of completeness and/or soundness. Indeed, we and others (Passemiers et al., 2022; Siahpirani & Roy, 2017) have consistently noted relatively low values of GRN reconstruction accuracy on real data benchmarks.

Our comparative evaluation involved four methods that have been found in a recent benchmarking study as leading methods for the task (Passemiers et al., 2022) and employ a variety of techniques (Random Forests, Gradient Boosting, Least Angle Regression, Precision Matrix calculation). Notably, only one of these tools (PORTIA) (Passemiers et al., 2022) is recent, while the others are over 10 years old. This reflects the reality that recent related work has not focused on the core problem of GRN inference from expression data, instead aiming to exploit additional types of data (e.g., epigenomic profiles and motifs) or tackling the noisy nature of single-cell data.

The performance of GRN inference methods generally does not depend on the number of target genes in the network, but it does depend on the number of candidate TFs and the number of actual regulators of a gene. The largest number of TFs considered in any of our tests is about 330 TFs, which is much smaller than the ~ 2000 TFs in human. However, in practice one expects GRN inference tools to be run with a selected subset of TFs, e.g., those with detectable expression or differential expression in the samples of interest. We recommend such pre-selection as it may help improve the accuracy of inferred GRN edges, and caution against using SPREd with thousands of TFs as candidate regulators.

Several avenues of future improvements to SPREd present themselves. Firstly, systematic tuning of hyperparameters of the neural network model should allow us to train more accurate models, especially if combined with even larger numbers of training samples, which are easy to generate. Secondly, we note that the current model uses as features five different types of pairwise relationships between genes, three of which (covariance, Pearson correlation, Precision matrix) capture linear relationships, one (Spearman correlation) captures monotonic relationships and one (mutual information) is more general. Our ablation analysis (Supplementary Figure 2) suggested that these five measures have mutual redundancies and it may be possible to remove one or more of these features. We leave this as a future modification, to be evaluated extensively in varying benchmark settings. Moreover, it will be interesting to explore if a 2D convolution of the joint distribution between two genes, akin to the approach of (Yuan & Bar-Joseph), can improve the featurization of pairwise relationships used by SPREd's neural network. We used a CNN architecture in SPREd as it is better able to suitably combine the five different

measures of pairwise expression relationships between TF-TF pairs and TF-gene pairs, compared to a simpler MultiLayer Perceptron model (Supplementary Table 5); however, future work may explore alternative and more effective architectures. Finally, given that SPREd does not perform parameter training, which is a strategy successfully used by its competing methods, a combination of these two complementary approaches will be a promising direction for future research.

GRN inference methods in the literature have generally been evaluated using a common benchmark called “DREAM5 GRN challenge” (Marbach et al., 2012). We chose the 3-layer GRNs and corresponding SERGIO-simulated expression datasets as our primary benchmark as we wanted our evaluations to systematically vary different parameters that control the difficulty of the problem. Note that the expression data simulator (SERGIO in “clean” mode) we used to generate the benchmark is very similar to the GeneNetWeaver (Schaffter et al., 2011) simulator used in DREAM. The main difference between the DREAM benchmarks and ours is in the underlying GRNs, which are far more numerous and varied in our benchmark.

CHAPTER 3. SINGLE-CELL TRANSCRIPTOMIC AND CHROMATIN DYNAMICS OF THE FLOW-INDUCED REPROGRAMMING OF ENDOTHELIAL CELLS IN ATHEROSCLEROSIS

3.1 Introduction

As one of the most devastating and prevalent cardiovascular diseases worldwide, atherosclerosis underlies the pathological process responsible for myocardial infarction, stroke, and peripheral artery disease (Benjamin et al., 2019; Herrington et al., 2016; Libby, 2021; Libby et al., 2009). Atherogenesis, the pathogenesis of atherosclerosis, encompasses a complex cascade of cellular and molecular events that culminate in the formation of atherosclerotic plaques within arterial walls. These events include extensive gene expression changes as well as major shifts in cell type and sub-type proportions in the affected tissues (Glass & Witztum, 2001; Tabas et al., 2015). Therefore, decoding the intricate gene regulatory networks (GRNs) in atherogenesis has become paramount for discovering novel biomarkers for early disease detection and developing targeted therapeutic interventions.

Atherogenesis involves a complex and dynamic interplay between various cell types, including vascular endothelial cells (ECs), smooth muscle cells (SMCs), and a diverse array of infiltrating immune cells, particularly macrophages (Kardassis et al., 2025). Central to the initiation and progression of atherosclerosis are ECs, which form the critical, mechanosensitive interface between blood flow and the arterial wall (Jebari-Benslaiman et al., 2022). While fundamental to vascular homeostasis, ECs are subjected to reprogramming by atherogenic stimuli of disturbed-flow (D-flow) along with

hypercholesterolemia (HighChol) (Chiu & Chien, 2011; Kumar et al., 2014; Lavillegrand et al., 2024; Park et al., 2025). This reprogramming drives ECs from an atheroprotective state toward several pro-atherogenic phenotypes, including chronic inflammation, endothelial-to-mesenchymal transition (EndMT), and endothelial-to-immune cell-like transition (EndIT) that we previously identified (Andueza et al., 2020). Flow-induced reprogramming of ECs (FIRE) (Tamargo et al., 2023) and endothelial dysfunction represent early detectable abnormalities in atherogenesis. However, the complex transcriptional mechanisms of FIRE remain to be fully elucidated.

For decades, our understanding of the molecular basis of atherogenesis has been largely informed by bulk RNA sequencing analyses (Romanoski et al., 2010). Although these approaches have established a critical framework of atherosclerosis, they are limited by the resolution of population-averaged data, masking molecular signals and the distinct contributions of the many different cell types and sub-cell types present within the heterogeneous plaque environment. This cellular heterogeneity is a critical feature of atherogenesis, as the dynamic shifts of cell abundance and phenotypic switches between subpopulations dictate the development of the disease (Fernandez et al., 2019; Wirka et al., 2019). The application of single-cell multi-omics in atherosclerosis research has revolutionized our ability to deconstruct this complexity (Williams et al., 2020). The integration of single-cell RNA sequencing (scRNA-seq) and chromatin accessibility profiling (scATAC-seq and ChIP-seq) analysis in the mouse partial carotid ligation (PCL) model with induced D-flow in the presence of hypercholesterolemia enables the characterization of EC heterogeneity during atherogenesis (Andueza et al., 2020; Dunn et al., 2014; Nam et al., 2009; Son et al., 2013), revealing the spectrum of subpopulations with specialized functions and distinct regulatory states. In the backdrop of these recent developments and our on-going work with the PCL model, our specific goal was to determine the key regulators of EC transcriptomic dynamics during atherogenesis.

There is a multitude of available computational strategies that have been proposed for recovering transcriptional regulatory relationships between transcription factors, TFs, and genes of interest, based on transcriptomics and/or epigenomics data (Aibar et al., 2017; Fleck et al., 2023; Pratapa et al., 2020). The transcriptomics data may reveal that a TF-gene pair have coordinately varying expression from cell to cell or across cell types, suggesting a regulatory relationship (Weidemuller et al., 2021). Complementarily, evidence of a TF's binding site in gene-linked cis-regulatory element, commonly revealed by epigenomics data, also supports such a TF-gene relationship (Pliner et al., 2018). Indeed, state-of-the-art tools for GRN reconstruction combine both of these lines of evidence for robust inference (Kamimoto et al., 2023; Moerman et al., 2019). Our approach builds on this established paradigm, with several key modifications, outlined below.

First, we focused our analysis exclusively on the genes that are differentially expressed between the different sub-types of endothelial cells (ECs), whose varying abundances have been found recently to be a key facet of D-flow induced atherogenesis (Andueza et al., 2020). This allowed our findings to narrow in on the specific regulatory mechanisms underlying EC dynamics during the process. Second, we quantified TF-gene co-expression patterns in ways that emphasize EC-associated transcriptomic differences. Third, while we included all major cell types in our expression models, we added cell types as covariates in the models to suppress the importance of regulatory relationships associated with cross-cell type variations and to focus more on intra-EC variation and co-variation. Fourth, we incorporated cis-regulatory evidence into GRN inference at two different levels of granularity: (i) TF binding site in a gene-linked CRE, and (ii) TF binding site overrepresentation in CREs linked to a set of EC-associated genes. The former provides more specific evidence but is typically less reliable, while the latter is a coarser form of evidence that is more statistically sound (Lareau et al., 2019).

Through comprehensive analysis of single-cell multi-omics data and development of machine learning-based pipelines, we identified a core set of transcription factors (TFs), Creb3l2, Rela, and Mef2c, that drive the FIRE in atherogenesis. These master regulators represent key nodes in the GRNs that govern the transition from atheroprotective to pro-atherogenic endothelial phenotypes, providing novel insights into the mechanistic basis of flow-mediated atherosclerosis. Notably, we showed experimentally that perturbing the expression levels of these core transcription factors modulates their predicted target genes, many of which are pathologically relevant to atherogenesis. This regulatory hierarchy demonstrates the central role of these transcription factors in controlling atherogenesis and validates their potential as therapeutic targets for preventing atherosclerotic cardiovascular diseases.

3.2 Methods

3.2.1 Animal studies

All animal procedures were approved by the Emory University Institutional Animal Care and Use Committee (IACUC) under protocol PROTO202100052 and were conducted in strict accordance with federal guidelines and regulations. The study utilized a total of 10 male C57BL/6 mice, 9 to 10 weeks old, obtained from Jackson Lab for single cell assay for transposase-accessible chromatin sequencing (scATAC-seq). Following the procedures performed for scRNA-seq data (Park et al., 2025), mice received a tail-vein injection of AAV8-PCSK9 (1x10 viral genomes; Vector Biolabs #AAV8-D377Y-mPCSK9) to achieve low-density lipoprotein receptor (LDLR) knockout and induce hypercholesterolemia. Concurrently, mice were fed a high-fat Western diet beginning one week prior to surgery

and maintained on it for the duration of the study. To induce disturbed flow (d-flow), mice underwent partial carotid ligation (PCL) surgery as previously described (Nam et al., 2009). Under anesthesia, three of the four caudal branches of the left common carotid artery (LCA)—the left external carotid, left internal carotid, and left occipital arteries—were ligated. The right common carotid arteries (RCAs) were used as contralateral controls (Nam et al., 2009). Exposure of the LCA to d-flow post-surgery was confirmed using ultrasound (Nam et al., 2009). Mice were sacrificed for scRNA-seq experiments by CO₂ asphyxiation at four weeks following the PCL surgery.

3.2.2 *Isolation of nuclei from carotid arteries for scATAC-seq assay*

Single-cell suspension preparation and sequencing were performed as previously described (Andueza et al., 2020; Park et al., 2025). Following euthanasia, blood was collected from the inferior vena cava using a heparin-coated 25G needle (Air-Tite #N2558) for plasma cholesterol analysis. Briefly, the left and right common carotid arteries (LCAs and RCAs) were cleaned, perfused with saline solution, and dissected. For an initial luminal digestion, the ends of the arteries were ligated, and a dissociation buffer was injected into the vessel. The buffer was composed of 500 U/mL collagenase type I (EMD Millipore #SCR103), 500 U/mL collagenase type II (MP Biomedical #0.210050.2.5), 150 U/mL collagenase type XI (Sigma-Aldrich #C7657), 60 U/mL hyaluronidase type I-S (Sigma-Aldrich #H3506), and 60 U/mL of DNASE I (Zymo #E1011) in HBSS (Cytiva #SH30031), filtered through a 0.45 µm syringe filter (Celltreat #229753). The arteries were incubated for 35 minutes at 37°C. After incubation, the buffer containing the lumenally-released cells was flushed into a 1.5 mL Eppendorf tube with fetal bovine serum (FBS) to neutralize enzymatic activity and placed on ice. The remaining arterial tissues were then

minced with microscissors in fresh dissociation buffer and incubated for an additional 45 minutes at 37°C. This second cell solution was also moved to a tube containing FBS on ice. Both the luminal and leftover digest samples were combined and filtered through a 70 µm cell strainer. The resulting cell suspension was washed with 2% bovine serum albumin (BSA) in PBS and centrifuged at 1000 g for 5 minutes. Red blood cells were lysed by incubating the cell pellet in Hybri-Max buffer (Sigma-Aldrich #R7757) for 5 minutes at room temperature. To ensure a single-cell solution, a final dissociation step was performed by incubating the cells in accutase (Sigma-Aldrich #A6964) for 5 minutes at 37°C and resuspended in 2% BSA solution. Then, as done previously (Andueza et al., 2020), the single-cell suspension was washed with 0.04 % BSA twice, followed by incubation with chilled lysis buffer composed of 10mM Tris-HCl, 10mM NaCl, 3 mM MgCl₂, 0.1% Tween-20, 0.1% NP40, 0.01% digitonin and 1% BSA in nuclease-free water for 5 min on ice. The cell lysates were washed with chilled wash buffer composed of 10 mM Tris-HCl, 10 mM NaCl, 3 mM MgCl₂, and 0.1% Tween-20 and resuspended in Nuclei Buffer (10X Genomics PN-2000153/2000207). Single nuclei encapsulation and library preparation were performed at the Emory Integrated Genomics Core (EIGC) using the 10X Genomics Chromium Next GEM Single Cell ATAC Kit v1.1 on the Chromium X device. The resulting cDNA libraries were sequenced on an Illumina NovaSeq instrument to a minimum depth of 25,000 reads per nucleus.

3.2.3 CRISPR interference TF validation

Telomerase-immortalized human aortic endothelial cells expressing doxycycline-inducible dCas9-KRAB-BFP (teloHAECs) for CRISPR interference (CRISPRi) were kindly provided by Drs. Rajat Gupta and Jesse Engreitz (Schnitzler et al., 2024). For each

TF (CREB3L2, MEF2C, RELA), top 2 paired sgRNA oligonucleotides targeting the promoter region were selected from the publicly available hCRISPRi-v2.1 library (Horlbeck et al., 2016) (Supplementary Table 7). A non-targeting (NT) guide sequence was chosen from the study (Doench et al., 2016). The pXPR_050 vector (Addgene 96925) (Sanson et al., 2018) was modified to replace the puromycin resistance gene with a blasticidin resistance gene derived from LentiGuide Blast (Addgene 199622). TF-targeting guides, as well as a non-targeting guide for control, were cloned into the modified pXPR_050-Blast vector by Golden Gate Assembly with BsmBI-v2. The plasmid constructs were verified by Sanger sequencing and packaged into lentiviruses using a standard three-plasmid protocol (Sanson et al., 2018) with minor modifications. Briefly, a second-generation lentiviral vector system was used. HEK293T cells (Takara Lenti-X 293T) were co-transfected with the generated CRISPRi plasmids (pXPR_050-Blast), pCMV_VSVG (Addgene 8454) envelope plasmid, psPAX2 (Addgene 12260) packaging plasmid using FuGENE HD (Promega) according to the manufacturer's instructions. Lentiviral-containing conditioned media were harvested 48 hours post transfection, centrifuged and filtered with 0.45 micrometer PES filter. The teloHAECs were transduced by incubation with lentivirus for 24 hours in the presence of 8ug/ml polybrene (Tocris), followed by blasticidin selection (15 ug/ml Invivogen) for 4 days. The transduced cells were then treated with 2 ug/ml doxycycline (Sigma Aldrich) for 4 days to activate the CRISPRi machinery in 100 mm dishes, with N=6 dishes for each group.

3.2.4 RNA Isolation, qPCR, and Bulk RNA-seq

The teloHAECs were lysed using the QIAzol lysis reagent (Qiagen #79306) (Tamargo et al., 2024) and total RNA preparation was performed using the Direct-zol RNA

Miniprep Kit (Zymo Research #R2052), including treatment with DNase I (Zymo Research E1012). After RNA concentration measurement using NanoDrop Spectrophotometer (Thermo Scientific 2000c). A fraction of the isolated RNA was saved for qPCR, while the leftover RNA was submitted to the Emory EIGC for bulk RNA-seq.

For qPCR, High Capacity cDNA Reverse Transcription Kit (Applied Biosystems #4368813) was used for cDNA synthesis and PerfeCTa SYBR Green FastMix (QuantaBio #95073–05K) was used for qPCR on the StepOnePlus Real-Time PCR System (Applied Biosystems #4376600). GAPDH was used as a loading control and quantifications were performed by $\Delta\Delta C_t$ method. Primers used for each target gene is included in Supplementary Table 8.

For bulk RNA-seq, EIGC performed QC on NanoDrop and Agilent BioAnalyzer 2100 to determine RNA concentration and RIN. Illumina NovaSeq X Plus Series was used for sequencing with a targeted depth of 50 M reads per sample.

3.2.5 Bulk RNA-seq Analysis

The *Rsubread* R package was used to build a genome index for hg38 version of the human genome for read alignment. The fastq files that contained sequencing outputs were aligned to the created human genome index. The reads were quantified for each gene using the BAM output files and converted to DGEList object using *edgeR* R package. Human *org.Hs.eg.db* R package was used for gene annotation, and only the genes with count-per-million (CPM) value greater than 0.19 in at least 2 out of 24 sample libraries were kept for QC, filtering out lowly expressed genes. The generated counts matrix with raw, unnormalized read counts for each gene in each sample and metadata file denoting the

sample information were used as inputs for *DESeq2* differential gene expression (DGE) analysis. The DGE analysis was conducted separately for each TF against NT control as a reference. Thereafter, normalization and model fitting for each TF vs NT control dataset were performed, and differential gene expression lists relative to NT control were generated for each TF. For visualization, volcano plots were generated using the *EnhancedVolcano* R package for the differentially expressed genes comparing each TF vs NT control.

3.2.6 *Analysis of differential gene expression between EC sub-types in scRNA-seq data*

We performed two types of differential expression analysis in EC subclusters present in the scRNA-seq data from (Park et al., 2025), using the FindMarkers function in the Seurat package (Hao et al., 2024). To identify marker genes in each EC sub-types for functional annotations, each EC subcluster was compared with all other EC sub-types. To identify genes that changed expression along the transition between sub-types of ECs, the atheroprotective EC subcluster (EC1) was compared with each of the subcluster of ECs from 2 through 5. The analysis was conducted using the default Wilcoxon rank sum test. A gene was considered significantly differentially expressed the adjusted p-value by Bonferroni correction was less than 0.06. To ensure the robustness of the comparisons, we only considered genes detected in a minimum of 10% of cells in either the EC1 or the other EC subclusters. This procedure yielded a list of genes that are significantly upregulated in EC1 (EC1_EC*) or downregulated (EC*_EC1) in EC1 relative to the other endothelial cell subtypes.

3.2.7 *scATAC-seq processing and analysis pipeline*

The processing of single-cell ATAC-seq (scATAC-seq) data was performed using the R packages Signac (v1.14.0) and Seurat (v5.3.0). The filtered peak-by-barcode count matrix for each sample including 4 week PCL PCSK9 LCA and RCA left over, 2 day PCL LCA and RCA, and 2 week PCL LCA and RCA, was loaded from the 10X Genomics CellRanger hdf5 file, along with the corresponding metadata. An initial quality control filtering was applied to retain only high-quality cells with more than 500 unique nuclear fragments. A Fragment object was then created from the fragments file for these cells. This information was used to construct a ChromatinAssay object, which was further filtered to include peaks present in a minimum of 10 cells per peak and cells containing at least 200 detected peaks. The ChromatinAssay object was then used to initialize a final Seurat object. The genomic annotations for the mm10 mouse genome were retrieved from the EnsDb.Mmusculus.v79 database and integrated into the Seurat object to facilitate the linking of peaks to genes. The data was normalized using the Term Frequency-Inverse Document Frequency (TF-IDF) method. The top features (peaks) were selected using a minimum cutoff of 20 cells, and dimensionality reduction was performed via Singular Value Decomposition (SVD) to generate Latent Semantic Indexing (LSI) components. A UMAP embedding was then generated for visualization using LSI dimensions 2 through 50. The processed Seurat object was then saved for subsequent analysis.

For 4 week experiments, the above processing was first performed for scATAC-seq data from both LCA and RCA samples. Upon manual inspection, RCA samples were excluded due to low-quality, resulting in a final cohort of 8,128 scATAC-seq nuclei for downstream analysis.

3.2.8 *scRNA-seq and scATAC-seq integration*

To integrate the scATAC-seq data with our comprehensively annotated scRNA-seq dataset, we employed the label transfer and data integration framework within Seurat v5. First, to create a comparable feature space between the two modalities, a gene activity matrix was calculated from the scATAC-seq data using the GeneActivity function. This process estimates gene expression by summing fragments within gene bodies and promoter regions, using the variable features from the scRNA-seq object as the gene set. This gene activity matrix was then added as a new assay to the scATAC-seq Seurat object, which was subsequently log-normalized and scaled. The cross-modality anchors were identified using the FindTransferAnchors function, with the scRNA-seq object serving as the reference and the scATAC-seq object's gene activity assay as the query. The integration was performed in a shared low-dimensional space derived from canonical correlation analysis (CCA). These anchors were first used to project the detailed cell type annotations from the scRNA-seq reference onto the scATAC-seq cells via the TransferData function, using the LSI dimensionality reduction from the scATAC-seq data to weight the transfer. Next, to enable joint visualization, the same anchor set was used to impute an RNA-like expression profile onto the scATAC-seq cells, which was stored as a new integrated assay. Finally, for co-embedding, unique cell identifiers were added, and the scRNA-seq reference object was merged with the annotated and imputed scATAC-seq object. This combined Seurat object was then scaled and subjected to Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) using the shared variable features. The resulting UMAP visualization (used in Figure 3.1) allows for cells from both modalities to be displayed in a single, integrated space, confirming successful alignment and consistent cell type annotation across the datasets.

3.2.9 *Chromatin co-accessibility and CRE identification*

For each cell type, we performed a co-accessibility analysis using Cicero (v1.9.2) within the Monocle3 (v1.3.1) framework. The filtered peak-by-cell matrix from the 10x Genomics output was used to create a cell data set (cds) object as the input for the Monocle3. Cells were filtered to retain those with a total number of accessible peaks between 2,000 and 15,000, and peaks with 0 reads were removed. Dimensionality reductions were performed using LSI based on SVD, followed by UMAP. A Cicero cds object that aggregated similar cells was then created by k-nearest-neighbors algorithm from the UMAP embedding. Next, the `run_cicero` function using the mm10 genomic coordinate as reference was applied to the Cicero cds object, yielding a list of co-accessibility scores between all pairs of peaks within a genomic distance of 500kb. The transcription start sites (TSSs) were annotated using mm10 as the reference genome for all peaks in the cds object and merged with Cicero output to derive a peak-to-gene list. After setting the co-accessibility score of TSS peak associated with the TSS site to 1, a threshold on the co-accessibility score of ≥ 0.8 was applied to filter peaks, returning a final list of peak-gene associations. Each such gene-linked peak was designated as a candidate cis-regulatory element (CRE) of the gene.

3.2.10 *TF-CRE and gene relationships*

To identify the transcription factors (TFs) that may bind to the gene-linked CREs identified above, we performed motif scanning using `gimmemotif` (van Heeringen & Veenstra, 2011) with motif database from `gimmemotifs motif v.5`. The `gimmemotifs` scanner searches the DNA sequence of each ATAC-seq peak region for matches to the

motifs in the PFM database. This scan generates a "TF-to-peak" information file, which served as the cis-evidence matrix. This file was structured as a dictionary or similar data object where each open chromatin peak was linked to a list of TFs that have predicted binding motifs within that peak.

This pipeline resulted two TF-gene maps representing predicted regulatory interactions between TFs and their target genes, one based on motifs and one based on ChIP-seq data on TF-binding.

3.2.11 Key TF identification in cis-regulatory analysis in EC subpopulation

Hypergeometric tests were performed to determine the statistical significance of the association between a TF and a set of differentially expressed genes (DEGs). This test evaluates whether the number of genes shared between the TF's target gene set (derived from motif- or ChIP-based approaches) and the DEG set of EC1 and other sub-types of ECs is greater than what would be expected by random chance. For each test, the parameters were defined as follows: (1) the total number of genes in the scRNA-seq dataset as the background universe; (2) the total number of genes in the EC1_EC* DEG set; (3) the total number of genes in the TF's target gene set based on motif or ChIP analysis; and (4) the number of genes in the intersection of the two sets. The resulting p-value represents the probability of observing an overlap of that size or greater by chance. A p-value < 0.001 was interpreted as evidence of a potential regulatory relationship between the TF and the biological process represented by the DEG set.

3.2.12 Differential accessibility and enrichment analysis

TF-CRE associations with co-accessibility scores of scATAC-seq data collected from LCA and RCA after 2 day, 2 week, 4 week PCL were derived as above respectively. The co-accessibility scores of the peaks of LCA from either 2 week or 4 week PCL (2w-l and 4w-l) were compared with the co-accessibility scores of the peaks of 2 day PCL RCA (2d-r) were compared after the peaks between each pair of data were intersected. The peak was defined as differentially accessible if the co-accessibility score changed from <0.8 in 2d-r to >0.8 in 2w-l or 4w-l, meaning that the peak became accessible after treatment, or from >0.8 in 2d-r to <0.8 in 2w-l or 4w-l, meaning that the peak became inaccessible, by which 4 differentially accessible (DAP) peak sets were derived. The enrichment tests were then performed between each of the of DAP sets and each of the EC1_EC* or EC*_EC1 DEG set using hypergeometric test described in the section above.

3.2.13 Gene expression prediction using random forest regressor

We implemented a Random Forest Regressor model using the scikit-learn library. A separate model was trained for each individual target gene. For each model, the dataset was partitioned into a training set (80%) and a held-out test set (20%). The Random Forest Regressor was then trained on the training data. The performance of each trained model was evaluated on both the training and test sets by calculating the R^2 .

3.2.14 Gene expression prediction using MLP

We developed a Multi-Layer Perceptron (MLP) regression model for each gene in any of the DEG sets, using the PyTorch deep learning framework. The input feature vector comprises gene expression levels of all TFs in a cell (sample), and the output is a target gene's expression in that cell. The model architecture consisted of a sequential network

that begins with a dropout layer ($p=0.9$) for regularization. This is followed by two hidden layers of 128 and 32 units, respectively, each using a Rectified Linear Unit (ReLU) activation function. The network culminates in a final linear layer with a single output neuron to predict the target gene's expression value. For each model, the dataset was split into training (80%) and testing (20%) sets. The models were trained to minimize the Mean Squared Error (MSE) loss between the predicted and actual expression values, using the Adam optimizer. To prevent overfitting, an early stopping mechanism was implemented, which terminated the training process if the training loss did not improve over 10 consecutive epochs. The final predictive performance of each model was assessed on the held-out test set by calculating R^2 . Integrative Gradients (IG) scores were used to interpret the trained models and quantify the importance of each input transcription factor. (We explored use of SHAP scores as an alternative, but chose IG instead due to run time considerations.)

3.3 Results

This project was conducted in collaboration with Christian Park and Dr. Hanjoon Jo (Emory University). The experimental data used in this study, including preparation of samples, generation and initial processing of the scRNA-seq dataset, and generation of ATAC-seq dataset, were carried out by our collaborators. Their contributions provided the foundational dataset and ensured high-quality input for the subsequent computational analyses.

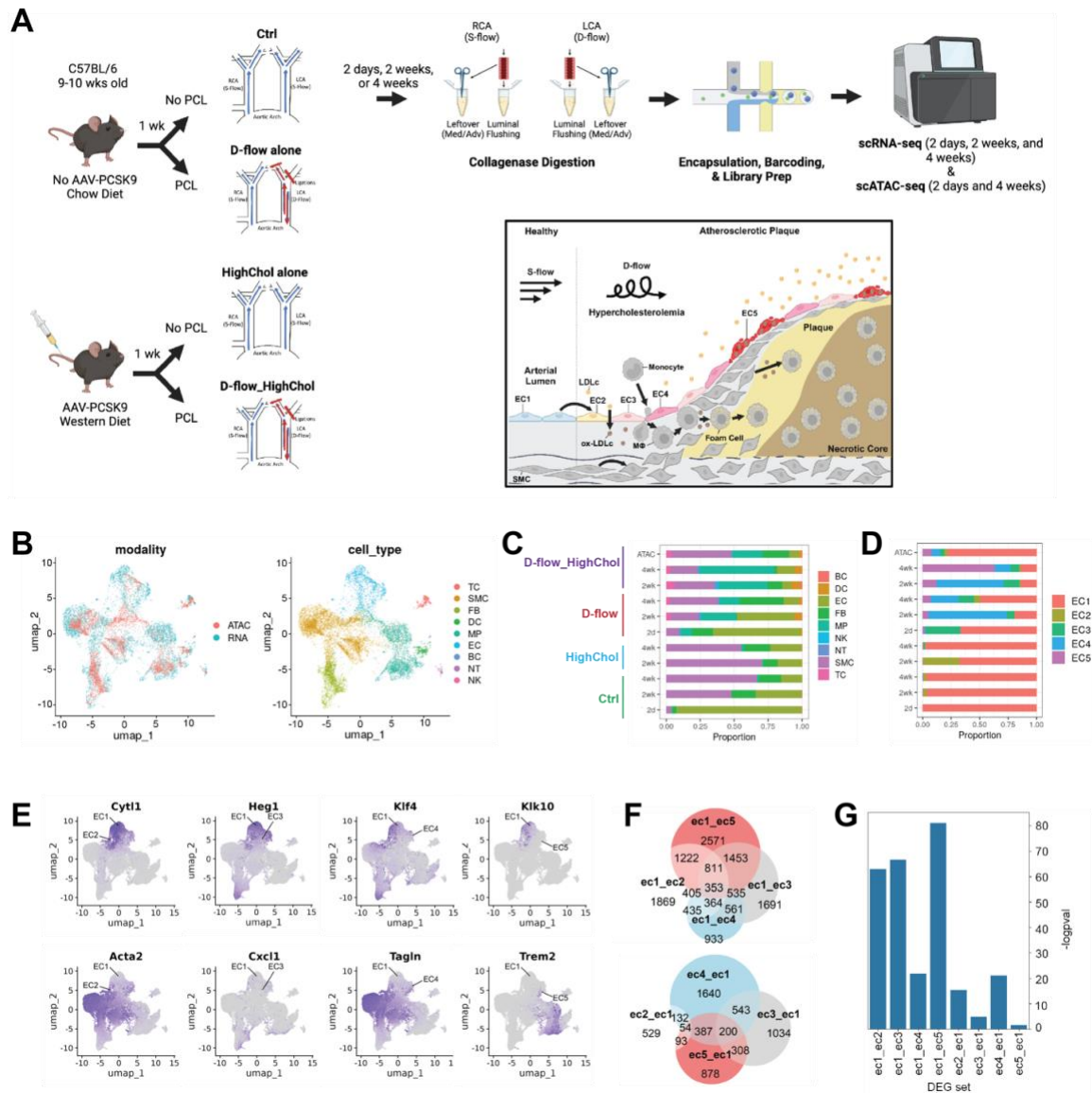


Figure 3.1 Single-cell multi-omics overview. (Some of the schematics were adapted from Park et al.)

(A) Schematic representation of atherosclerosis development and experiment designs. C57BL/6 mice (N=5-20 mice) were treated with or without AAV-PCSK9 and Western diet to induce hypercholesterolemia and with or without PCL surgery on the LCA for its exposure to disturbed flow (d-flow) a week later. After 2 or 4 weeks, the mouse carotid arteries were isolated and enriched with endothelial cells by luminal flushing, followed by further digestion of the leftover arteries. The flushing and leftover samples were separately processed into single cells, encapsulated, barcoded, and sequenced for scRNA-seq analysis. Furthermore, 2 days and 2 weeks samples containing only the luminal digestion samples (N=10 mice) were added for a complete analysis (Andueza 2020), and scATAC-seq was performed on LCAs of additional

C57BL/6 mice (N=10 mice) treated with AAV-PCSK9 and Western diet, followed by PCL surgery a week later. (B) UMAP plot of co-embedded scRNA-seq and scATAC-seq data (each point is a cell), colored by modality (scRNA-seq or scATAC-seq) and by cell type. (The scRNA-seq data comprised ~10x more cells than scATAC-seq, hence we downsampled scRNA-seq cells for ease of visualization.) (C) Cell type proportions of cells in scATAC-seq data from D-flow, high cholesterol treatment at 4 weeks (top row) and in scRNA-seq data from different treatments and time points (remaining rows). (D) Endothelial cell sub-type (EC1-EC5) proportions in scATAC-seq (top row) and scRNA-seq (remaining rows) in the same experimental conditions as in panel C. (E) Examples of differentially expressed genes (DEGs) between the atheroprotective “EC1” sub-type of endothelial cells and each of the other four sub-types (EC2, EC3, EC4, EC5). Each panel shows a gene’s expression (gene name shown at top) as a feature map on the common UMAP plot of all cells, with clusters of the two contrasted sub-types labeled. The top row of panels illustrates genes overexpressed in EC1 while the bottom row illustrates genes downregulated in EC1. (F) Venn diagram showing the sizes and overlaps of gene sets differentially expressed between EC1 and each of the other four sub-types of EC cells. A gene set labeled “EC1-EC*” represents genes overexpressed in EC1 compared to another sub-type, while a label of “EC*-EC1” represents genes down-regulated in EC1 relative to the other sub-type. (G) Bar plots showing $-\log_{10}$ p-value of Hypergeometric tests of intersection between differentially expressed gene sets (between EC1 and each of the other four sub-types of Ecs) and genes linked to differentially accessible peaks between LCA and RCA at 2 days.

3.3.1 Single-cell multi-omic analysis in atherogenesis

Our goal was to delineate the cellular and regulatory landscape during atherosclerosis progression by capturing both transcriptomic dynamics and chromatin accessibility profiles (Figure 3.1A). The experimental design included mice subjected to partial carotid ligation (PCL) surgery on the left carotid arteries, fed either a standard chow or a Western diet. We adopted scRNA-seq data from the right and left carotid arteries (RCA and LCA) at 3 key time points - 2 days, 2 weeks, and 4 weeks post-PCL (Methods). In parallel, we performed scATAC-seq on LCA samples from hyperlipidemic mice at 4 weeks (Methods). We also utilized previously published scATAC-seq data from the PCL mouse model (Andueza et al., 2020). This dataset comprised samples collected at 2 days and 2 weeks post-PCL from both the LCA and RCA. We report results from a total of 98,553

cells for scRNA-seq, 18,324 cells for 2 days and 2 weeks post-PCL scATAC-seq, and 8,128 nuclei for 4 weeks post-PCL scATAC-seq after preprocessing on each dataset (Methods).

We integrated the scRNA-seq data with scATAC-seq data (4 weeks post-PCL) into a shared latent space using established methodology (Methods), in order to assign cell types to all cells in a consistent manner and also to correct for technical differences (batch effects) between the two modalities. Upon integration, the scATAC-seq cells were annotated by label transfer from pre-annotated scRNA-seq cells (Park et al., 2025). Visualizing the integrated data using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018), we noted well-defined cell clusters (Figure 3.1B) with each cluster comprising a mix of RNA-seq and ATAC-seq cells, confirming a successful integration while preserving distinct representations of different cell types. We then compared the relative abundance of each cell population in both scATAC-seq and scRNA-seq datasets under the condition of disturbed-flow (D-flow), hypercholesterolemia (HighChol), both disturbed-flow and hypercholesterolemia (D-flow_HighChol), and neither (Ctrl) (**Figure 1C**). We found the cell type proportions under conditions of D-flow_HighChol had a strong concordance between the scATAC-seq and scRNA-seq data. Specifically, the fraction of endothelial cells (EC), smooth muscle cells (SMC), and T cells (TC) in scATAC-seq closely matched the proportions in scRNA-seq. Important to our cis-regulatory analysis, the major cell types in RNA-seq data -- endothelial cells (EC), smooth muscle cells (SMC), fibroblasts (FB) and macrophages (MP) – were well represented in the ATAC-seq data. This allowed us to define accessibility landscapes of each of these cell types in the D-flow_HighChol condition, for use in our cis-regulatory analyses below.

3.3.2 *Reprogramming of endothelial cells by D-flow under hypercholesterolemia*

We then examined the heterogeneity within the subpopulation of endothelial cells (Figure 3.1D), since these sub-types have been previously noted as associated with atherogenesis (Park et al., 2025). Five major sub-types of ECs have been identified and functionally characterized based on their marker – “EC1” as an atheroprotective sub-type, “EC2” as contractile endothelial-to-mesenchymal transition (EndMT), EC3 as proinflammatory cells, EC4 as synthetic, and EC5 as endothelial-to-immune cell like transition (EndIT) and endothelial-to-foam cell transition (EndFT) cells. We observed shifts in the proportions of these EC sub-types in the RNA-seq data across the four conditions, e.g., there is a clear decrease in EC1 sub-type in the disturbed flow conditions, while the EC3-EC5 sub-types are only seen in those conditions, suggesting a pathological transition in the endothelial cells under the pressure of disturbed-flow. These observations, previously reported in (Park et al., 2025), prompted a detailed examination of the molecular changes associated with the shift from the atheroprotective EC1 sub-type to the other sub-types, including identification of differentially expressed genes and their regulatory mechanisms. Notably, all five EC sub-types were represented in the ATAC seq data, which is important for identifying cis-regulatory regions associated with their transcriptomic signatures.

We systematically investigated the upregulated and downregulated genes in EC1 (the atheroprotective sub-type) versus each of the other four EC sub-types (EC2-EC5). We identified sets of DEGs (Bonferroni corrected p-value < 0.05, Methods) for each of the eight contrasts. As expected, these DEG sets exhibited high degree of mutual overlap when comparing the EC1-upregulated DEG sets among each other and, and similarly for the

EC1-downregulated gene sets (Figure 3.1F). Examples of DEGs from each of the above eight sets are visualized in Figure 1E, showing higher color density (expression level) in one cluster (sub-type) of cells compared to another. Complementing the differential expression analysis, we analyzed differentially accessible chromatin regions associated with disturbed flow-induced changes. We identified non-coding genomic segments whose DNA accessibility, measured by scATAC-seq, differs between post-PCL LCA (representing disturbed flow) and post-PCL RCA (representing control condition). We then identified the genes linked to these differentially accessible chromatin regions (Methods), and tested these genes for enrichment in the above-identified DEGs, for evidence of cis-regulatory dynamics underlying the extensive transcriptomic response noted above. This analysis was performed separately with LCA accessibility data from three different time points (2 days, 2 weeks, 4 weeks), in each case comparing with RCA data from 2 days as control. The strongest enrichments of DEGs were obtained for differentially accessible regions between LCA and RCA at 2 days (Figure 3.1G), implying an early epigenetic switch triggered by pro-atherogenic conditions. The EC1-EC* gene sets, i.e., genes up-regulated in EC1 (atheroprotective sub-type), implying an early epigenetic switch triggered by pro-atherogenic conditions. (See Supplementary figure 9 for corresponding results from 2 weeks and 4 weeks LCA.) The EC1-EC* gene sets, i.e., genes up-regulated in the atheroprotective EC1 sub-type, were especially enriched for differential accessibility in their cis-regulatory regions, suggesting that the global chromatin changes associated with atherogenesis have, on average, a repressive effect on atheroprotective genes' expression.

3.3.3 Cis-regulatory analysis reveals transcription factors underlying endothelial cell reprogramming

We next sought to identify transcription factors (TFs) that regulate the transcriptomic changes between EC sub-types reported above. We adopted a two-pronged approach to this, taking inspiration from prior work (Kamimoto et al., 2023) (Fleck et al., 2023)– first we examined cis-regulatory regions of genes for evidence of TF binding, a hallmark of TF-gene regulatory relationship, and second, we examined evidence for TF-gene expression covariation across cells, as is expected if a TF regulates the gene. We first report the results of our cis-regulatory analysis.

We used scATAC-seq data to identify “peaks” of high DNA accessibility, which serve as a comprehensive set of potential cis-regulatory elements (CREs). We linked these CREs to genes based on co-accessibility patterns using the Cicero tool (Pliner et al., 2018) (Methods). Gene-CRE relationships were assigned in a cell type-specific manner, i.e., each gene was assigned its putative CREs in a cell type using ATAC-seq profiles of cells of that type. We focused on the relationships for endothelial cells (ECs). Next, we scanned the CREs for TF binding sites using either computational scanning with motifs (Methods) or catalogs of genome-wide binding profiles (ChIP-seq) from the GTRD database (Kolmykov et al., 2021) (Methods). Combining these TF-CRE relationships with the CRE-gene links established above, we obtained cis-regulatory maps of TF-gene associations (Methods) reflecting evidence of a TF binding at an accessible CRE associated with the gene. Two such maps were obtained, one based on motif scanning (1095 TFs) and the other on ChIP-seq data (745 TFs). We note that each map has limitations and is expected to be noisy – the motif-based map is the result of a simplistic computational prediction pipeline that does not account for the context-specificity of TF-DNA binding, while the ChIP-based map utilizes experimental data on TF binding in diverse cellular contexts, most of which differ

from the tissue and conditions analyzed here. Our rationale was that either of these provides evidence of TF binding *potential* and combining it with CRE evidence from ATAC-seq data incorporates information about cellular context, resulting in a useful cis-regulatory map.

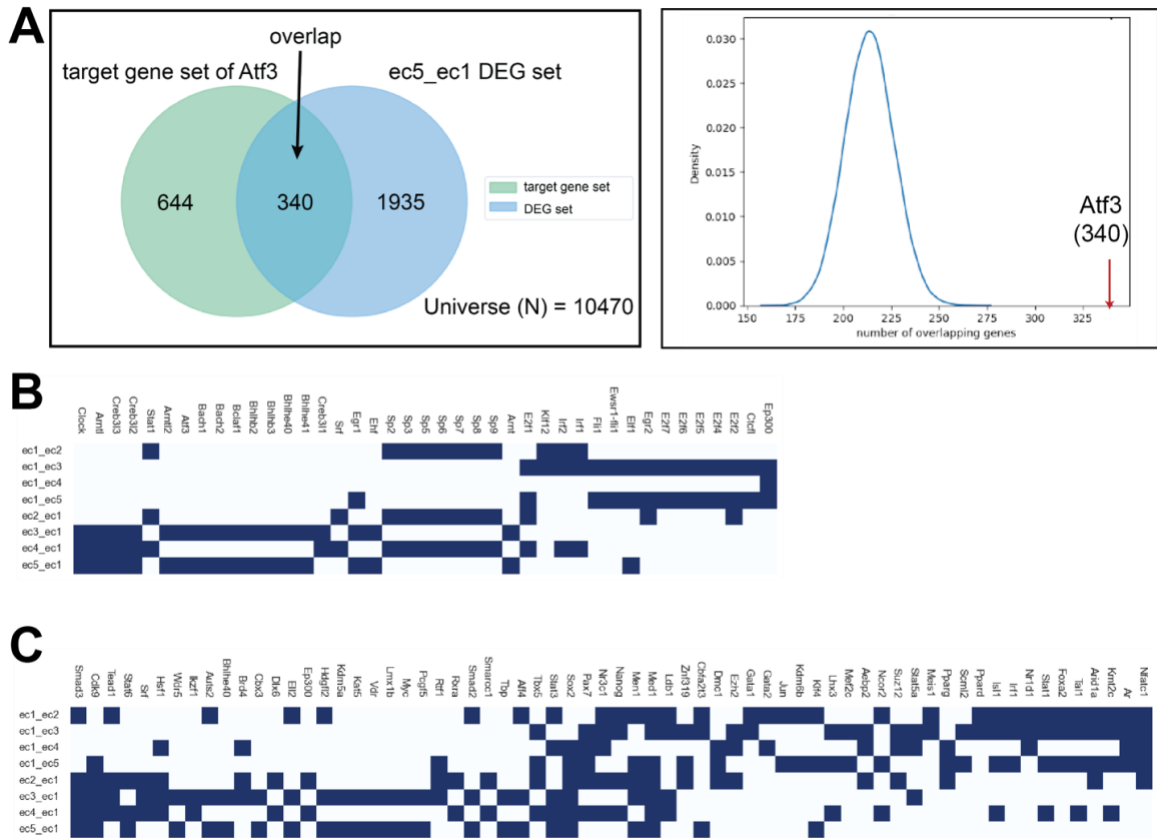


Figure 3.2 Cis-regulatory analysis of differentially expressed genes (DEGs) between endothelial cell (EC) sub-types associated with atherogenesis. (A) Illustration of how associations between TFs and DEG sets are discovered. Venn diagram showing the overlap between a DEG set (EC5-EC1) and the target gene of a TF (ATF3) reveals a substantial overlap, which is found to be statistically significant (p-value < 0.001 for motif-based analysis and p-value < 0.0001 for ChIP-based analysis) by a Hypergeometric test. On the right is shown the null hypergeometric distribution and the value of the observed overlap size (red arrow) is clearly significant. (B, C) Most significant TFs associated with each DEG set based on motif analysis (B) or ChIP analysis (C).

We then connected the above cis-regulatory maps to endothelial cell reprogramming by testing for statistical association between each TF's regulon (its target genes according to the map) and each DEG set (genes differentially expressed between the atheroprotective EC1 sub-type and other EC sub-types). The association between the target gene set of a given TF and a specific EC DEG set was quantified by a hypergeometric test that evaluates the significance of the overlap between the two gene sets, as illustrated for the TF ATF3 and the "EC5_EC1" DEG set in Figure 3.2A. Applying this statistical test for all TFs and all DEG sets, we identified significant 41 TF-DEG set associations using the motif-based cis-regulatory map (p -value < 0.001) (Figure 3.2B) and 66 associations using the ChIP-based map (p -value < 0.0001) (Figure 3.2C).

Our findings from both the motif- and ChIP-based analysis were supported by previous reports. For example, KLF2 is a well-established atheroprotective factor: laminar shear stress robustly induces its expression, which in turn sustains the production of endothelial nitric oxide synthase and thrombomodulin, thus maintaining a quiescent endothelial phenotype even under inflammatory stimuli (van Thienen et al., 2006). KLF4, another KLF family member, suppresses proliferation in vascular smooth muscle cells and supports anti-inflammatory, anti-thrombotic signaling, thereby mitigating atherogenic remodeling (Tetreault et al., 2013). ELF1, an ETS transcription factor, has a critical role in vascular gene regulation—particularly of the vessel-specific *Tie2* gene during development—highlighting its contribution to endothelial integrity and possibly vascular disease states (Huang et al., 2006). GATA2 drives pro-atherogenic behavior in endothelial and immune cells; it enhances monocyte adhesion and vascular permeability via upregulation of adhesion molecules, and its overexpression in macrophages disrupts

efferocytosis, a key mechanism in early plaque formation (Luo et al., 2022). Finally, BHLHE40, a circadian-clock linked bHLH factor, shows elevated serum levels in individuals with subclinical atherosclerosis, positively correlating with carotid intima-media thickness—suggesting a potential biomarker or active participant in vascular pathology (Xu et al., 2023).

These results highlighted distinct sets of TFs predicted to regulate the gene expression changes that characterize the shift from EC1 sub-type to the more atherogenic sub-types EC2-5. We also noted that TF-DEG set associations obtained using motif-based and CHIP-based maps showed substantial degree of complementarity (Supplementary figure 10), with some degree of agreement as well. This led us to require both types of associations as cis-regulatory evidence when shortlisting the most important TFs in our integrative analysis below (Figure 3.4).

3.3.4 Explainable machine learning framework for TF discovery in atherogenesis

A complementary strategy commonly employed (Langfelder & Horvath, 2008) for discovering important TFs of a regulatory system is to infer TF-gene relationships based on expression co-variation across cells (Wu & Sinha, 2024), and identify TFs with large numbers of predicted target genes as being the most important. This approach is based solely on expression data and provides evidence of a TFs regulatory importance complementary to the cis-regulatory evidence pursued in the previous section. We adopted this approach (Huynh-Thu & Geurts, 2018; Kamimoto et al., 2023), implementing a machine learning (ML)-based method to infer TF-gene relationships from expression data. As outlined in Figure 3.3A, we first applied MAGIC imputation (van Dijk et al., 2018) to

address technical dropouts in scRNA-seq data, and normalized the data to prepare for ML modeling. Next, we constructed separate ML models for each gene that predict its expression as a function of the expression levels of all candidate TFs. We then used Integrated Gradients (IG) (Sundararajan et al., 2017), a popular feature attribution method, to interpret the trained model for a gene, thus assigning an importance score to each TF for that gene. These steps were performed for each gene that is differentially expressed between EC1 and any other EC sub-type, thus providing us a putative gene regulatory network (GRN) underlying all DEG sets.

We performed a series of rigorous evaluations to determine the most suitable ML model for our approach. We compared two leading and simple ML algorithms – Multi-Layer Perceptron (MLP, a simple neural network) and Random Forest (RF) and found MLP to demonstrate superior or comparable accuracy to RF in predicting the expression of target genes (Figure 3.3B). (Accuracy was assessed using R² on a “test set” of cells hidden during training.) We chose these two ML models because the RF underlies one of the most popular GRN prediction algorithms, GENIE3 (Huynh-Thu & Geurts, 2018), and our recent work found MLP to be an equally or more powerful approach. Based on our evaluations, we selected to use MLP for our GRN inference strategy. It is generally believed that a typical gene is under the regulatory influence of only a handful of TFs. To incorporate this common wisdom into the ML model, we incorporated 90% dropout in the input layer of the MLP, as a regularization strategy. Our evaluations showed that this modification greatly improves predictive accuracy with a mean test R² of 0.9625 compared to 0.8920 without the modification (Figure 3.3C). The resulting optimized MLP

architecture was therefore used for all subsequent GRN inference, providing a validated framework for our biological discovery.

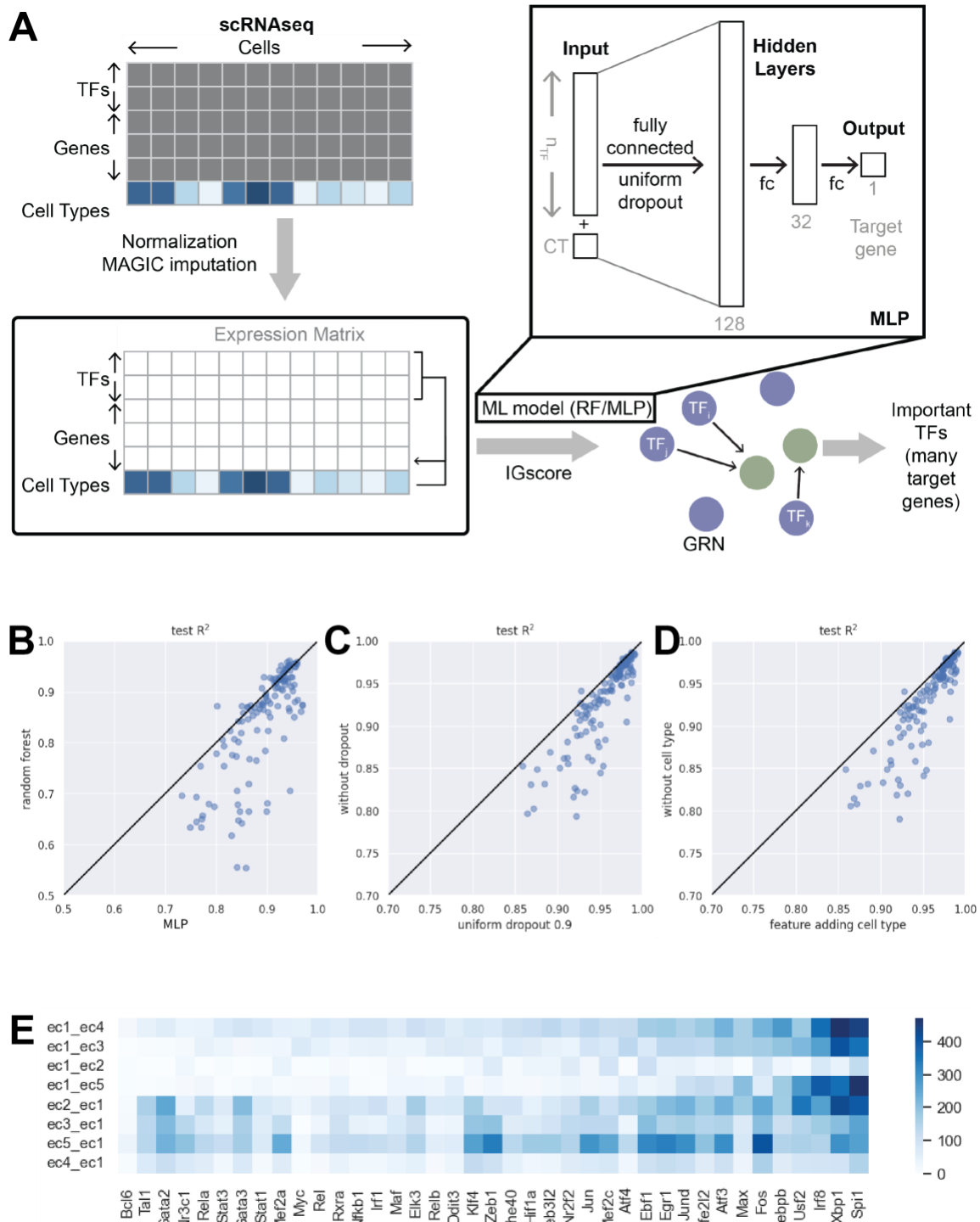


Figure 3.3 Expression-based discovery of TF-gene regulatory relationships. (A) Schematic of computational workflow. The input comprises scRNA-seq data, shown here as a matrix with rows being genes and columns representing cells. The input also includes cell type annotations of cells and a subset of genes (rows) marked as TFs (candidate regulators). The expression data are subjected to normalization and

imputation (using MAGIC) to obtain a processed expression matrix for further analysis, where each non-TF gene is modeled as a function of all TF genes using either Random Forest (RF) or MultiLayer Perceptron (MLP) regression. The cell type is also used as a covariate in this ML model. The trained ML model is interpreted by using “Integrated Gradients” (IG) scores that assign an importance value to each TF, reflecting the TF’s regulatory influence on the gene, which yields the gene regulatory network (GRN) edges leading into the gene. This is repeated for each non-TF gene. TFs with the largest number of target genes identified in this manner are considered the “important TFs”. The MLP architecture used is illustrated in the panel on top right. The input layer includes the expression level of all (n_{TF}) TFs and cell type (CT) of a cell, and the output is the target gene’s expression in that cell. Separate models are trained for each target gene. (B-D) Evaluation of ML models for predicting gene expression. Test R2 values are shown for a collection of XX target genes (DEGs of EC1 vs EC2-5), comparing (B) MLP model to Random Forest, (C) MLP model with a 90% dropout rate with or without cell type (CT) as a covariate in addition to all TFs, (D) MLP model with a 90% dropout rate in the first fully connected layer versus without dropout. (E) Important TFs for each DEG set. Shown is the number of target genes of a TF among genes of each DEG set, as determined from the GRN inference outlined in panel A. Top 10 TFs from each DEG set were chosen based on the top rank of the number of regulating edges and then were concatenated across DEG sets.

It is well known that cell types differ in their transcriptomic profile (Siletti et al., 2023). This implies that a significant portion of TF-gene co-expression results from the cell type-related differences in transcriptome. In other words, cell type is a potent confounder for statistical discovery of TF-gene relationships (such as the procedure above), and some prior studies have sought to address this by performing GRN inference for each cell type separately. Here we sought to tackle this confounding factor by using cell type as an additional covariate in the ML model. This allows us to train models on all cells rather than a much smaller population of cells belonging to a specific cell type, and at the same time avoids reporting spurious TF-gene relationships arising from the transcriptomic differences from one cell type to another. We repeated our benchmarking procedure with the MLP model with and without cell type information (as covariate) and confirmed that including cell type improves the test R2 of the model (Figure 3.3D).

Based on the above evaluations, we trained MLP models (with cell type as covariate and with dropout regularization mentioned above) to score the putative regulatory influence of TFs for each gene. We shortlisted the top 10% most important TFs (by IG score) for each gene, and counted, for each DEG set, the number of genes of that set that a TF targets. This count was used to score the overall regulatory importance of a TF to the gene set, yielding key regulators for each cellular state. The resulting TF-DEG set map (Figure 3.3E) includes several TFs with known function in endothelial cell function and/or atherosclerosis. For instance, it reports the TF Klf4 as having a high regulatory importance for genes down-regulated in the atheroprotective EC1 sub-type, consistent with previous findings (Park et al., 2025). On the other hand, the pro-inflammatory TF Atf3 was

identified among the top-ranking regulators for genes upregulated in the pathogenic EC5 sub-type (Park et al., 2025).

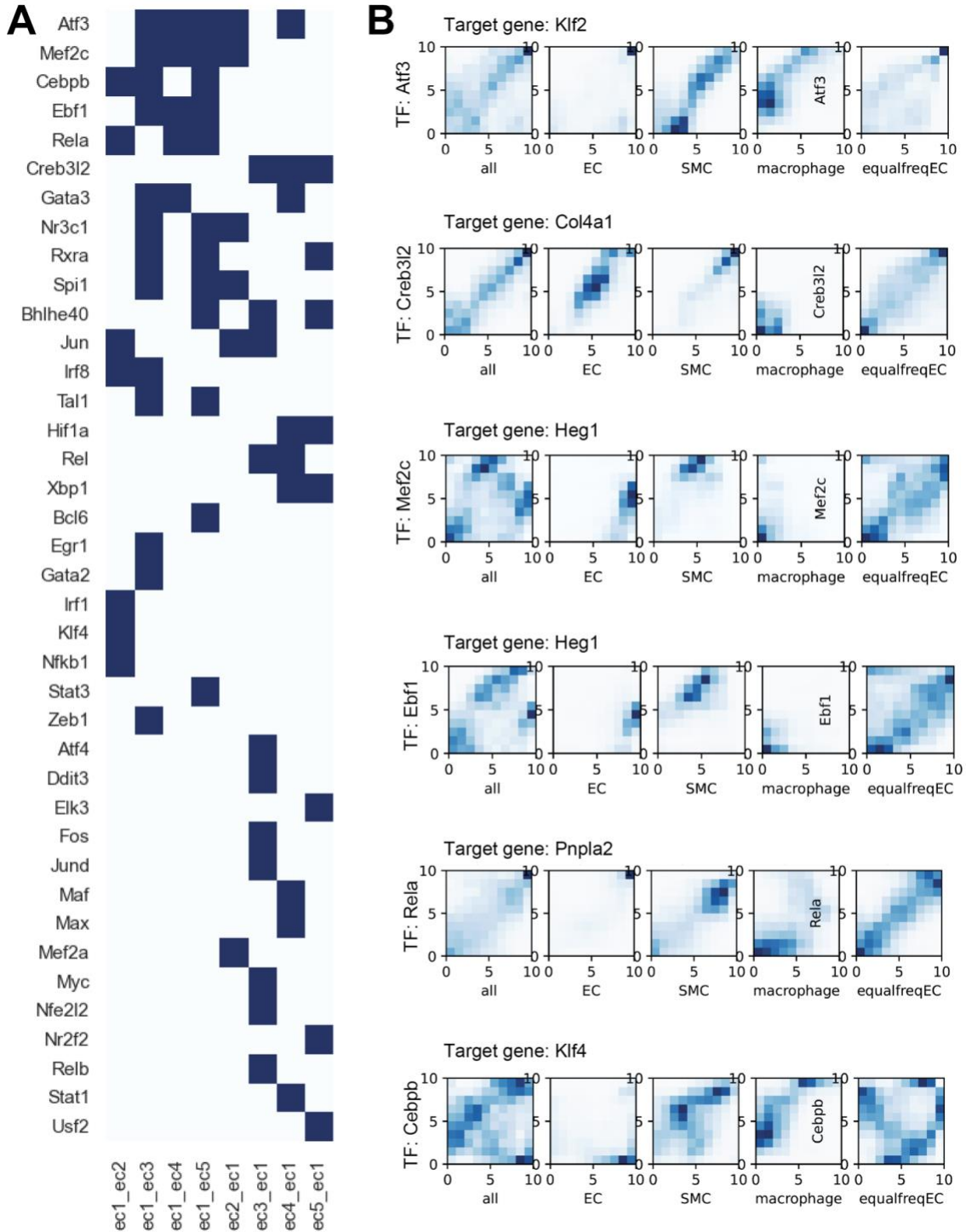


Figure 3.4 (A) Top TFs derived for each DEG set obtained by combining evidence from cis-regulatory analysis and expression-based analysis. Each TF is marked (blue cells) as important to a DEG set if the TF-DEG set association is statistically significant under both motif-based and ChIP-based analyses and expression-based analysis finds the TF to be among the top 10 by number of targets in the DEG set.

TFs are sorted by the row sum of the one-hot matrix then sorted by difference between the sum of EC1_EC* group and EC*_EC1 group. (B) Visualization of expression covariation between TFs and their predicted target genes, through six selected examples. Each row of panels represents a TF (one of the top six rows panel A) and one of its predicted targets (prediction supported by motif-based and ChIP-based analysis as well as expression-based analysis). The TF name is indicated on the left and the target gene name is on the top. Each panel is a 2D histogram of the TF's and target gene's joint expression distribution across a population of cells. Axes represent expression values scaled to a 0-10 range for each gene or TF. In each row, the left-most panel shows the 2D histogram across cells of all cell types, while the second, third and fourth panels show the histograms for endothelial cells (EC), smooth muscle cells (SMC) and macrophages respectively, the three most common cell types in the data. These four panels use the same "bins" of expression values. The fifth panel ("equalfreqEC") shows the same histogram as the second, i.e., for endothelial cells, but the "bins" of expression are set to be deciles of expression among endothelial cells, for easier visualization of TF-gene expression co-variation in this sub-population of cells.

3.3.5 Combining cis-regulatory and coexpression-based evidence provides a high-confidence set of TFs associated with endothelial cell reprogramming in atherogenesis

Our analyses above adopted two complementary strategies, one based on cis-regulatory evidence and the other based on expression covariation, to identify TFs associated with endothelial cell reprogramming by D-flow under hypercholesterolemia. We next combined the results of these analyses to obtain a high-confidence shortlist of regulators that are supported by cis-regulatory evidence from the motif-based as well as ChIP-based maps and are also among the top TFs identified by the ML-based expression analysis. The resulting TFs includes many examples of TFs associated with two or more DEG sets (Figure 3.4A).

Consistent with our findings, several TFs—including ATF3, CREB3L2, RELA, MEF2C, EBF1, and CEBPB—exhibit well-supported roles in regulating atherosclerotic

processes. ATF3 acts as a protective regulator in atherosclerosis, where its deficiency (e.g., *Atf3*^{-/-} in *ApoE*^{-/-} mice) enhances macrophage foam cell formation and plaque burden, while in human coronary lesions ATF3 correlates with plaque stability and reduced inflammatory markers like VCAM-1 and MMP-9 (Gold et al., 2012; Peng et al., 2021). CREB3L2 was identified among key regulators of *Vcam1*⁺ smooth muscle cell states in atherosclerosis, suggesting a role in modulating SMC phenotype transitions associated with lesion progression (Ord et al., 2023). RELA (NF- κ B p65) is a central effector of vascular inflammation, activating pro-inflammatory gene expression in endothelial, smooth muscle, and immune cells to drive lesion formation (de Winther et al., 2005; Tardajos Ayllon et al., 2025). MEF2C serves as an anti-inflammatory safeguard in endothelial cells by suppressing NF- κ B-mediated adhesion and cytokine signaling, thus helping maintain vascular homeostasis (inferred from its general regulatory context). EBF1 emerges through epigenetic studies as hypomethylated and potentially over-expressed in atherosclerotic plaque, linking this B-cell TF to disease pathogenesis (Yamada et al., 2014). Finally, CEBPB, part of the C/EBP family, is enriched in vascular smooth muscle and plaque-associated gene programs and contributes to inflammatory and fibrotic remodeling in atherosclerosis (Ord et al., 2023).

Figure 3.4B provides visual depictions of a select set of TF-gene relationships predicted by our pipeline. Each row of panels corresponds to a TF that is among the six key regulators noted above, and shows how the TF's expression covaries with an example target gene identified by the ML model. The selected target gene (different for each row) is a DEG between EC1 and one of the other EC sub-types, and harbors putative binding sites for the TF according to the motif-based as well as ChIP-based cis-regulatory maps. In

each row (depicting a TF-gene pair), the left-most panel shows the joint distribution of a TF (y axis) and its target gene (x axis) across all cell types, including EC, smooth muscle cell (SMC), and macrophage, along with the Pearson correlation coefficient (PCC) and corresponding p-value. We noted that in some cases the TF-gene relationship is roughly linear (e.g., ATF3-Klf2, CREB3L2-Col4a1, RELA-Pnpla2), while for others the relationship is more complex. Hypothesizing that the covariation patterns may differ between cell types, we then plotted analogous histograms for EC, SMC and macrophage cell types separately (remaining panels in each row). We observed the TF-gene covariation to be roughly linear in each case but in different ranges of TF expression. For example, while the overall covariation between the TF MEF2C and its predicted target gene Heg1 appeared to be nonlinear in the entire cell population, the expression in the EC sub-population of cells showed a strong positive correlation. In some cases, e.g., ATF3-Klf2 in macrophages, MEF2C-Heg1 in ECs and EBF1-Heg1 in ECs, the relationship appears to be non-linear even within a cell type. These results not only provide a tangible depiction of the GRN inferences made by our pipeline, they also validate our choice of using cell type as an additional covariate and the use of non-linear ML models in expression analysis.

3.3.6 Functional Validation of TFs reveals regulation of pro-atherogenic gene programs

To experimentally validate the function of top candidate TFs identified from our integrative analyses, we selected a panel of six high-confidence regulators: CEBPB, CREB3L2, ATF3, EBF1, MEF2C, and RELA (top six in **Error! Reference source not found.A**). We used a CRISPR interference (CRISPRi) system to individually knock down each of these TFs in endothelial cells and subsequently assess their impact on atherogenesis-relevant gene expression. The efficacy of each knockdown was first

quantified by qPCR. Of the six candidate TFs, we confirmed robust and statistically significant suppression of mRNA levels for CREB3L2, MEF2C, and RELA when compared to a non-targeting control (**Error! Reference source not found.A**). As the knockdown for CEBPB, ATF3, and EBF1 did not meet the pre-defined efficiency criteria (p -values > 0.05), these TFs were excluded from further investigation and all subsequent functional analyses were focused on the three TFs with successfully validated knockdown. Following validation, we assessed the transcriptomic consequences of depleting each TF using bulk RNA sequencing. This analysis revealed widespread changes in gene expression for each knockdown condition, as illustrated in the volcano plots showing numerous statistically significant DEGs compared to control cells (**Error! Reference source not found.B**). Crucially, a focused analysis of these DEGs from bulk RNA-seq data showed that the knockdown of CREB3L2, MEF2C, or RELA resulted in the significant downregulation of a key suite of pro-atherogenic marker genes (**Error! Reference source not found.C**). This targeted suppression of pro-atherogenic marker genes compared to EC functional genes and flow-induced reprogramming of endothelial cells (FIRE) genes supports that these TFs are key functional drivers of pro-atherogenic gene expression in endothelial cells. Thus, these experiments functionally validate our computational predictions, establishing CREB3L2, MEF2C, and RELA as critical regulators in endothelial cells during atherogenesis.

3.3.7 Global visualization of transcriptional regulatory profiles of genes

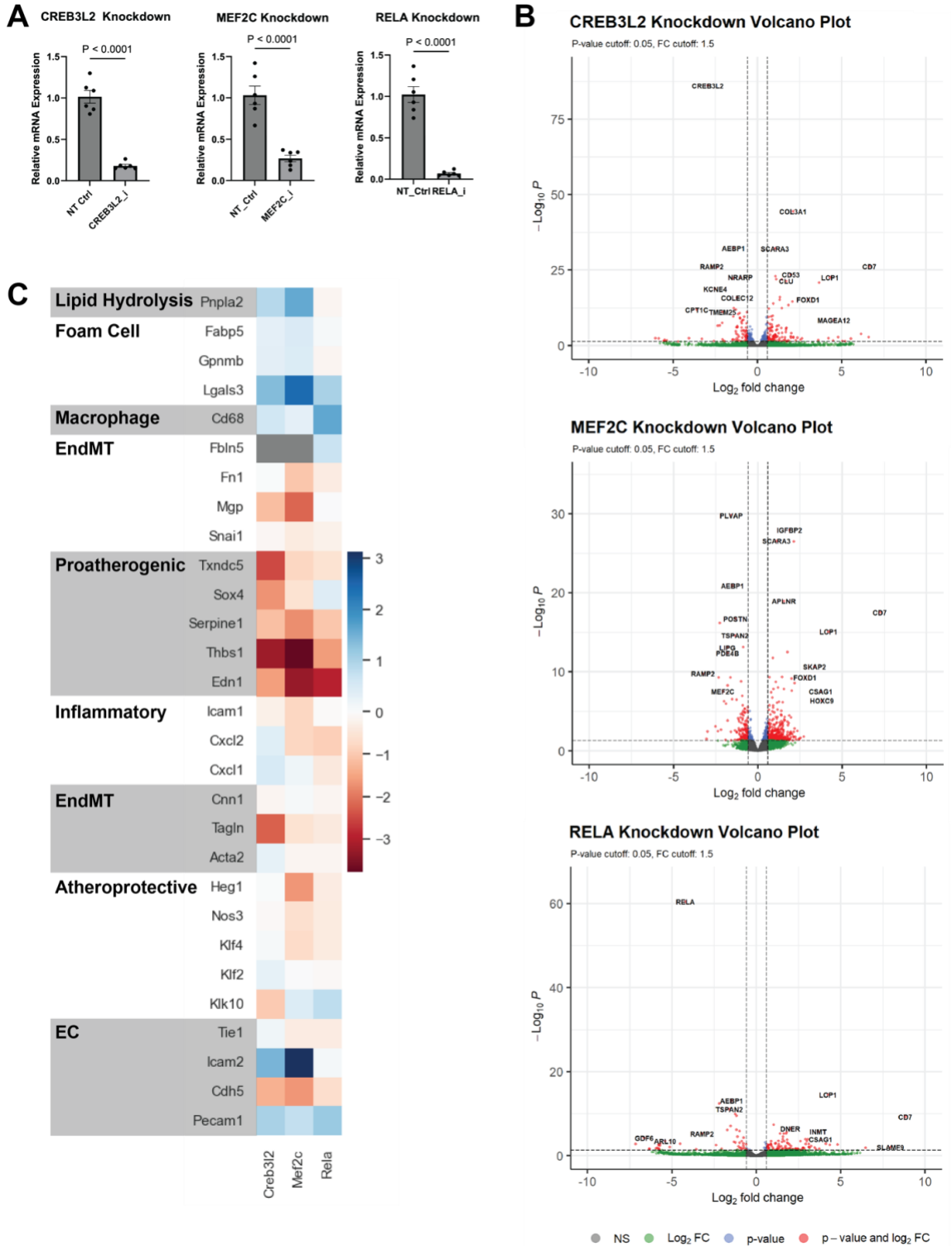


Figure 3.5 Experiment validation of top TFs predicted by cis-regulatory and expression-based analysis. (A) Results of qPCR validating successful CRISPR-mediated knockdown of CREB3L2, MEF2C, and RELA, three of the top six TFs

identified in the integrative analysis (Figure 3.4A), compared to the non-targeting control. (B) Volcano plots showing differentially expressed genes from bulkRNAseq comparing the CREB3L2, MEF2C, and RELA CRISPRi-knockdown cells and non-targeting control cells. Colors indicate if a gene is significantly DE at p -value < 0.05 and/or has a \log_2 fold-change $> 1e-6$ in absolute value, or neither (NS). (C) Heatmap on the left shows the \log fold-change in expression upon knockdown of TF compared to control cells, for key groups of genes identified previously. Proatherogenic marker genes were significantly down-regulated after knockdown of CREB3L2, MEF2C, and RELA, while such strong down-regulation was not observed for EC functional and proatherogenic FIRE marker genes. (Warm colors represent down-regulation on TF knockdown.) Heatmap on the right shows \log fold-change in expression of the same genes between the EC1 subtype and each of the more proatherogenic sub-types (EC2, ... EC5). (Warm colors represent up-regulation in the EC1 subtype.)

Our GRN modeling above, based on cis-regulatory evidence (motifs, ChIP-seq) and TF-gene co-expression, not only points us to key TFs of atherogenesis (experimentally validated above), it also provides a regulatory profile of each gene in the form of TFs predicted to regulate it. Specifically, a gene's regulatory profile comprises the three types of evidence (motif-based, ChIP-seq based and co-expression-based) for each TF's influence on that gene, i.e., a vector of $3 \times \#TFs$ dimensions, where $\#TFs=290$. We sought to visualize these high-dimensional gene-specific regulatory profiles for all DEGs, similar to how high-dimensional gene expression profiles of individual cells are visualized commonly. To this end, we devised a novel technique to reduce the high dimensionality of gene regulatory profiles while incorporating the expression readout of those profiles.

We designed a multi-label convolutional neural network classifier to simultaneously predict a gene's membership to each of the eight EC DEG sets (EC1_EC*, EC*_EC1) using its 290-dimensional regulatory profile (Figure 3.6A). After training the model, we extracted the 128-dimensional embedding of each gene (last hidden layer of neural network) and visualized these embeddings through t-distributed Stochastic Neighbor Embedding (t-SNE). As shown in Figure 3.6B, genes that are up- or down-

regulated in EC1 relative to another EC sub-type form distinct clusters in the t-SNE space, indicating that the learnt embeddings capture gene regulatory profiles in a biologically meaningful way.

3.3.8 Neural network model combines co-expression and cis-regulatory evidence for GRN inference

Our results above strongly suggest that combining motif- or ChIP-based evidence with findings from ML modeling of TF-gene co-expression provide a reliable means to inference of regulatory relationships. This is, in fact, consistent with existing strategies for GRN inference from single-cell data (Fleck et al., 2023; Kamimoto et al., 2023). However, in our method presented above, as well as in some of the leading existing methods (Huynh-Thu & Geurts, 2018; Moerman et al., 2019), the two lines of evidence are examined separately and then combined heuristically. Here, we investigated a more tightly integrated ML model that combines the evidence sources in a more principled manner. Specifically, if cis-regulatory analysis finds a TF to be significantly associated with a DEG set, we assigned that TF greater importance when training the ML model of gene expression.

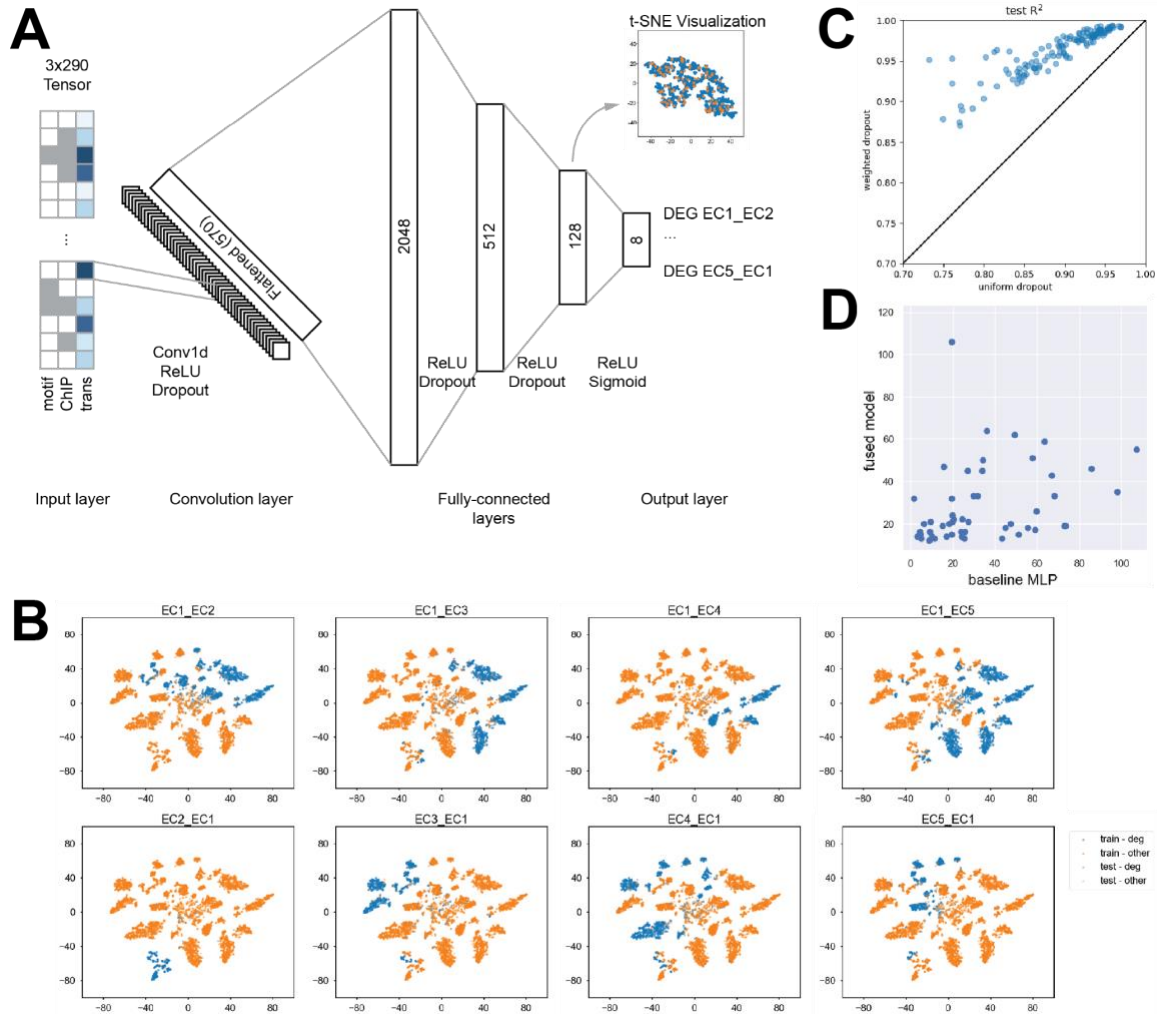


Figure 3.6 Enhanced prediction of gene regulation and classification of DEG sets through modality fusion. (A) Schematic of a classifier for EC DEG sets. The architecture of the neural network-based classifier was designed to predict a gene’s membership in each of eight DEG sets (DE between EC1 sub-type and other EC sub-types). The classifier takes results from motif-, ChIP-seq, and expression-based analysis as input and learns to categorize the input into EC1_EC* and EC*_EC1. The input feature vector has three dimensions (one for each type of evidence) for each TF, for a total of 3 x #TFs dimensions. (B) t-SNE plots showing embeddings of each gene obtained from the last hidden layer of the neural network of (A). Each panel shows the genes of a DEG set (EC1_EC*, EC*_EC1) in blue. Each vertically aligned pair of panels shows genes up- (top) and down- (bottom) regulated in EC1 compared to another EC subtype. (C) Comparison of accuracy of the baseline MLP model with uniform dropout and the modified MLP (fused model) that uses cis-regulatory evidence to inform dropout rates for the input layer. Shown are the test R² values for 0.96 randomly selected genes, representing accuracy of predicting gene expression across cells in an unseen random subset of cells, based on TF expression profiles of those cells. (D) Scatter plot that compares occurrence of top-ranked important TFs.

Each data point represented a TF with its count of as the top 10 regulators of a target gene in weighted dropout model (y-axis) and uniform dropout model (x-axis).

We modified the MLP model introduced above (for predicting a target gene's expression in a cell from the expression level of all TFs in that cell) to use prior importance scores of TFs derived from cis-regulatory analysis. This importance score was derived from the p-value of association between the TF and the DEG set that the target gene belongs to (see Methods). Recall that the MLP of Figure 3.4 uses dropout (= 90%) for its input layer as a means for regularization. We now set the dropout rate of each input unit (TF) to be proportional to its cis-regulatory importance score. Thus, with this minor modification, the MLP learns to prefer candidate TFs that have strong cis-regulatory association with the DEG set when modeling any gene in that set. This “weighted dropout” scheme is akin to use of priors in probabilistic models (Siahipirani & Roy, 2017). On the other hand, our strategy differs from prior work in that the cis-regulatory evidence used as prior information is at the level of TF-DEG set associations (a coarser but more reliable form of evidence) rather than TF-gene associations (more fine-grained but noisier information).

We compared the performance of the modified MLP (called “fused model”) against the exclusively expression-based MLP model that uses a uniform dropout rate. The comparison showed that the fused model with weighted dropout achieved improved R2 on the test set (Figure 3.6C), demonstrating a superior ability to accurately predict gene expression. This analysis indicates that weighting the contribution of different inputs (TFs) based on prior knowledge of cis regulation is an effective strategy for improving model performance. Given the enhanced predictive accuracy of the fused model, we next sought to determine if it could yield complementary biological insights compared to the baseline

MLP model. We compared the top-ranked important TFs identified by both the fused model and the baseline MLP, and observed a reordering of the most influential TFs identified by the model (Figure 3.6D). This analysis suggests that the fusion of multiple modalities was capable of reprioritizing regulatory relationships compared to single-modality models.

3.4 Discussion

In this work, we report findings of important regulators, *Creb3l2*, *Rela*, and *Mef2c*, of flow-induced reprogramming of endothelial cells (FIRE) during atherogenesis and present a framework in gene regulatory network (GRN) inference by extending existing approaches and integrating multi-omics single-cell data. The integration of scRNA-seq and scATAC-seq data through an explainable deep learning algorithm addresses a critical gap in current computational biology methodologies, where most existing approaches either focus on single-omics data or lack the interpretability necessary for biological insight generation.

The incorporation of cis-regulatory priors into the deep learning model constitutes a particularly noteworthy methodological innovation. By leveraging chromatin accessibility information alongside gene expression data, this approach moves beyond correlation-based inference to capture mechanistic regulatory relationships. The explainable nature of the model enhances interpretability of expression modeling, suggesting the potential mechanisms underlying these regulatory relationships. We also developed a new method of DEG set classifier, which converts features from both cis-evidence and expression-based models into a unified gene-embedding space. This method provides new insights on how

to meaningfully combine disparate data types while preserving their unique biological information content. The learned embedding space through classification enables visualization of differentially expressed gene sets for exploring complex regulatory landscapes.

The identification of key transcriptional regulators that drive atherogenesis represents a significant biological contribution with potential clinical implications. The ability to pinpoint central regulatory nodes in the disease process provides valuable targets for therapeutic intervention and biomarker development. Moreover, the single-cell resolution of the analysis enables the identification of cell-type-specific regulatory mechanisms, which is crucial for understanding the heterogeneous nature of atherosclerotic disease progression.

Despite its significant contributions, this work reveals several important limitations that highlight opportunities for future research. The observation that target gene sets of predicted key transcription factors showed enrichment only in athero-protective genes after experimental knockdown suggests an incomplete understanding of pro-atherogenic regulatory mechanisms. This asymmetry in regulatory understanding represents a critical gap that requires further investigation, as therapeutic interventions targeting pro-atherogenic pathways may be more clinically relevant than those enhancing protective mechanisms.

The need for systematic comparison with other complementary approaches to GRN inference from multi-omics data represents another important limitation. While the proposed method demonstrates clear advantages in terms of interpretability and biological

insight generation, a comprehensive benchmarking study would provide valuable context for understanding when and how this approach should be applied relative to alternative methods. Such comparative analysis would also help identify the specific strengths and weaknesses of the approach, guiding future methodological improvements.

The accuracy of individual gene targets identified for central transcription factors requires systematic assessment and validation. While the identification of key regulatory nodes is valuable, the precision of downstream target prediction directly impacts the utility of the inferred networks for hypothesis generation and experimental design. This limitation suggests the need for improved methods for target gene identification, possibly incorporating additional data types or more sophisticated modeling approaches.

Another limitation is the absence of spatial transcriptomics data, which could provide crucial contextual information on tissue-specific regulatory patterns. This limitation is particularly relevant in the context of atherosclerosis, a complex disease characterized by distinct spatial organization within arterial walls. The disease involves multiple cell types arranged in specific spatial patterns, with regulatory networks that likely vary significantly across different tissue regions and disease stages. The integration of spatial resolution with multi-omics modalities from the same cells would substantially improve both the accuracy and biological interpretability of inferred networks. Spatial transcriptomics could reveal how regulatory networks vary across tissue architecture, identify spatially restricted regulatory modules, and provide insights into cell-cell communication patterns that influence disease progression. This represents a particularly promising direction for future research, as spatial multi-omics technologies continue to mature and become more widely accessible.

Looking forward, this work establishes a strong foundation for continued advancement in multi-omics GRN inference. The emphasis on explainability and biological interpretability, combined with the demonstrated success in identifying disease-relevant regulatory mechanisms, provides a roadmap for applying similar approaches to other complex diseases and biological systems. The methodological framework developed here, while specifically applied to atherosclerosis research, has broad applicability and represents a significant step forward in our ability to understand complex biological systems through integrated multi-omics analysis.

The ultimate goal of such research—to understand the regulatory mechanisms underlying human disease and identify therapeutic targets—remains challenging but increasingly achievable through continued methodological innovation and the integration of emerging technologies. This work contributes meaningfully to that goal while clearly articulating the next steps necessary for continued progress in the field.

CHAPTER 4. CONCLUSION

This thesis explains a combination of computational approaches that can be used, in conjunction with existing tools, to infer gene regulatory network (GRN) in diverse scenarios marked by different types of data access and specific biological goals. Chapter 2 pursued a novel strategy to reconstructing TF-gene regulatory relationships from bulk transcriptomics data, without requiring additional data modalities, and we demonstrated its advantages over established machine learning-based approaches in scenarios where the number of distinct conditions represented in the data are relatively limited. Chapter 3 presented an end-to-end computational pipeline, rather than a single algorithm for a specific task, to analyze single cell transcriptomics and epigenomics data in contrasting conditions, with the specific goal of identifying TFs that drive atherogenesis. We were successful in identifying three such TFs, which will be the subject of future studies that detail their regulatory roles.

While each project offers valuable advances, there are also important limitations that highlight opportunities for future work. The SPREd algorithm of Chapter 2 trains a neural network model on synthetic data generated from a biophysics-based simulation model, alleviating the bottleneck of limited ground truth (training data) in many biological systems. However, several limitations still remain. First, the current model is designed exclusively for transcriptomics data, and does not accommodate multi-omics information such as chromatin accessibility or protein-DNA interactions, which can further reduce false positive regulatory edges by disambiguating direct from indirect regulation. Second, the framework does not yet incorporate single-cell-specific features such as adding technical

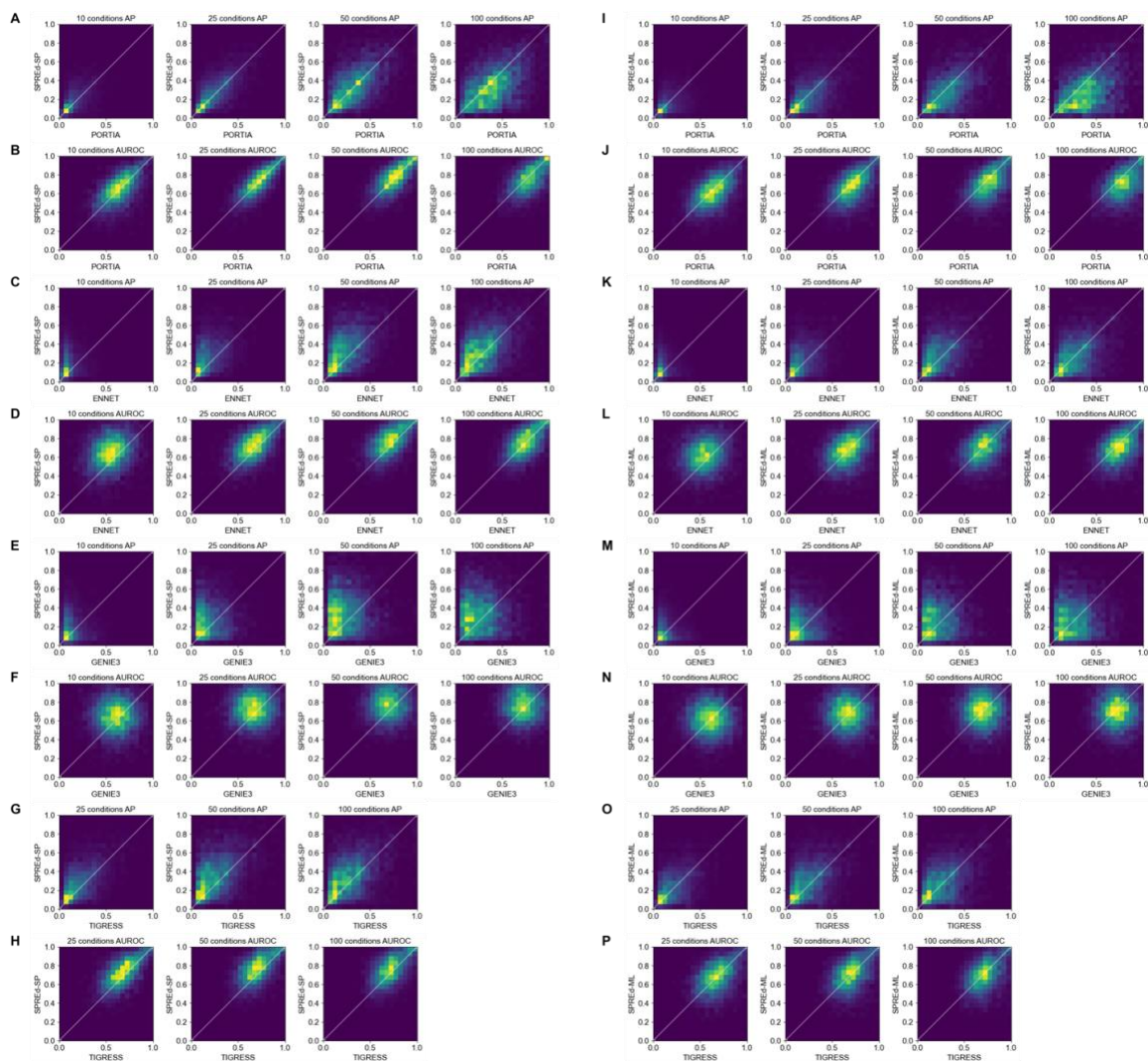
noise when generating the simulated data, limiting its resolution in highly heterogeneous systems. Third, we believe further improvements could be achieved through architecture design of the deep learning model and hyperparameter tuning, with the integration of other modalities. More importantly, the availability of real GRN benchmarks remains a critical limitation for evaluating and validating inferred networks.

The project presented in Chapter 3 extends GRN inference into the multi-omics, single-cell domain, specifically in the context of endothelial cell regulation during atherogenesis. By combining scRNA-seq and scATAC-seq data and incorporating cis-regulatory priors into an explainable deep learning model, this approach improves interpretability of expression modeling and identifies a set of key regulators that drive atherogenesis. It also provides a novel view of multi-modality fusion that converts features from cis-evidence and expression-based models into unified gene-embedding space, from which differentially expressed gene sets can be classified and visualized. This work also presents several opportunities for future improvement. As we found the target gene sets of the predicted key TFs showed enrichment only in athero-protective genes after experimental knockdown, there is a need for further investigations on the regulatory mechanisms of pro-atherogenic genes. Secondly, the results of our approach need to be examined in the light of other complementary approaches to GRN inference from multi-omics data, with a systematic breakdown of strengths and weaknesses of our approach. Thirdly, while we were able to identify TFs that play central roles in the reconstructed GRNs, the accuracy of individual gene targets of those TFs needs to be systematically assessed and improved methods may need to be developed for such identification. Another limitation is the absence of spatial transcriptomics data, which could provide valuable

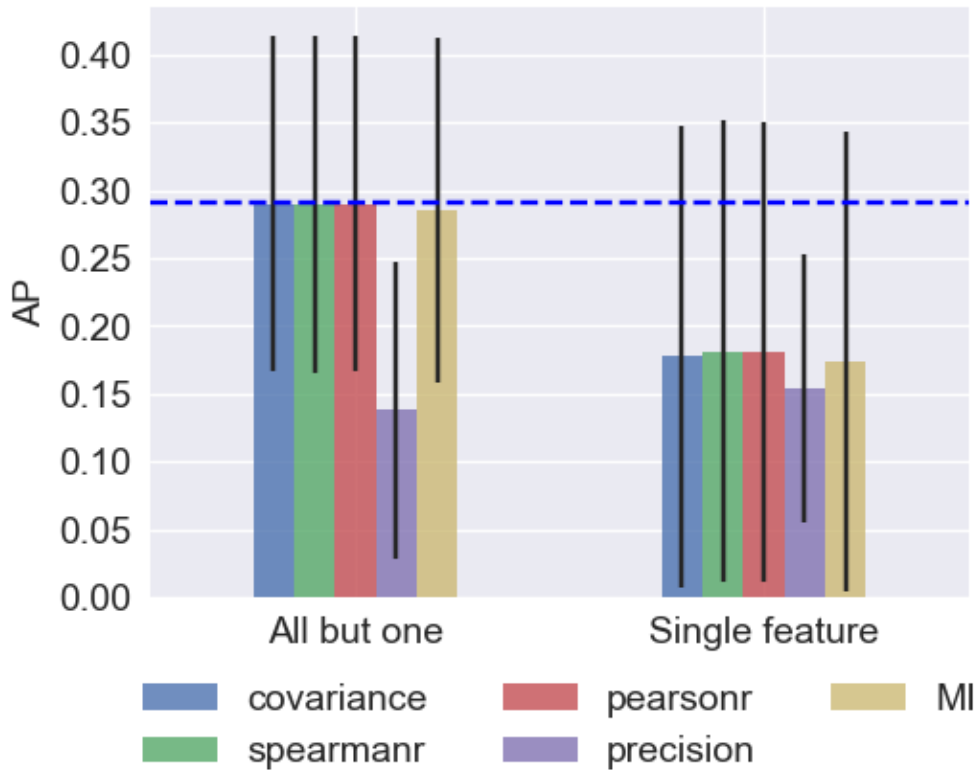
contextual information on tissue-specific regulatory patterns—particularly relevant in complex, spatially organized diseases such as atherosclerosis. Integrating spatial resolution and pairing multi-omics modalities from the same cells would substantially improve the accuracy and biological interpretability of inferred networks.

Together, the projects undertaken in this thesis illustrate the power of combining computational methods with domain-specific biological insight for GRN inference. Moving forward, integrating multi-omics data at single-cell resolution, incorporating spatial information, and validating predictions through perturbation experiments will be essential steps. Additionally, expanding the availability and quality of benchmark datasets will be critical for future method development and evaluation.

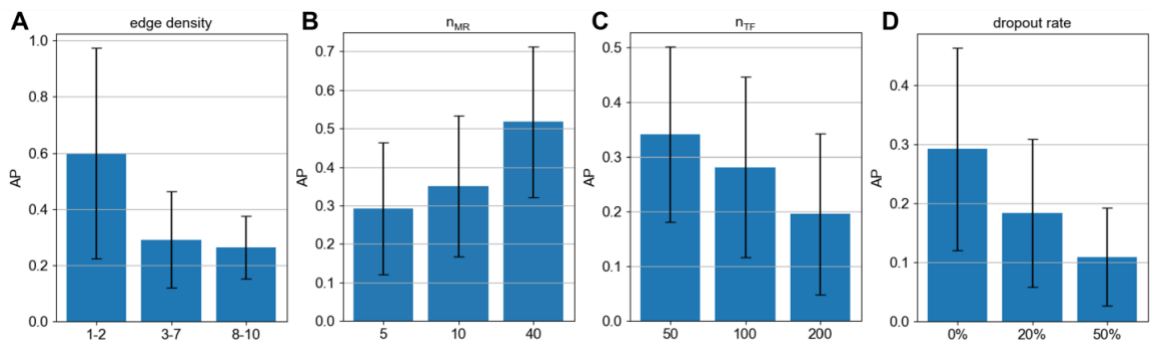
APPENDIX A. SUPPLYMENTARY FIGURES



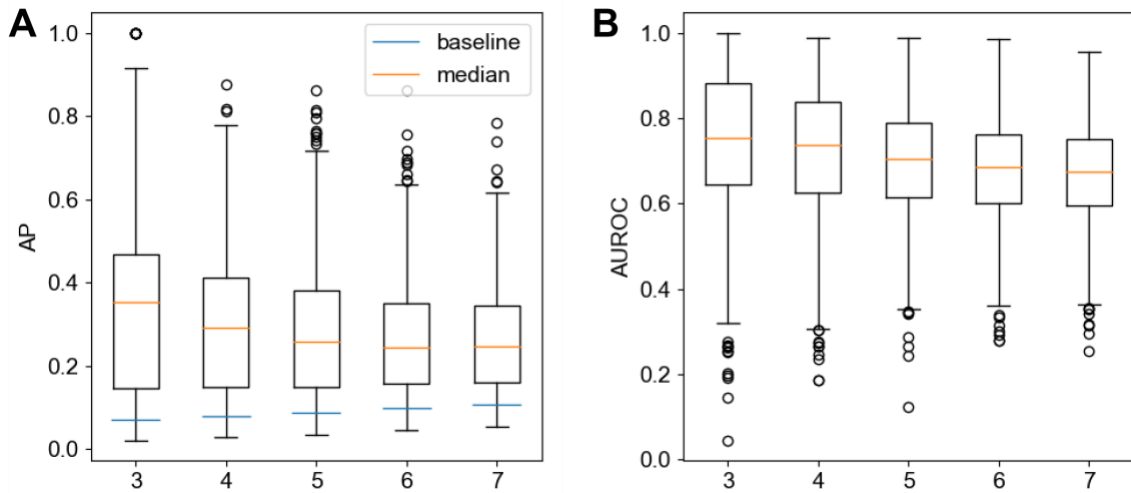
Supplementary Figure 1. Direct comparisons of AP (A, C, E, G, I, K, M, O) and AUROC (B, D, F, H, J, L, N, P) between SPRED-SP and PORTIA (A, B), SPRED-SP and ENNET (C, D), SPRED-SP and GENIE3 (E, F), SPRED-SP and TIGRESS (G, H), SPRED-ML and PORTIA (I, J), SPRED-ML and ENNET (K, L), SPRED-ML and GENIE3, or between SPRED-ML and TIGRESS (O, P).



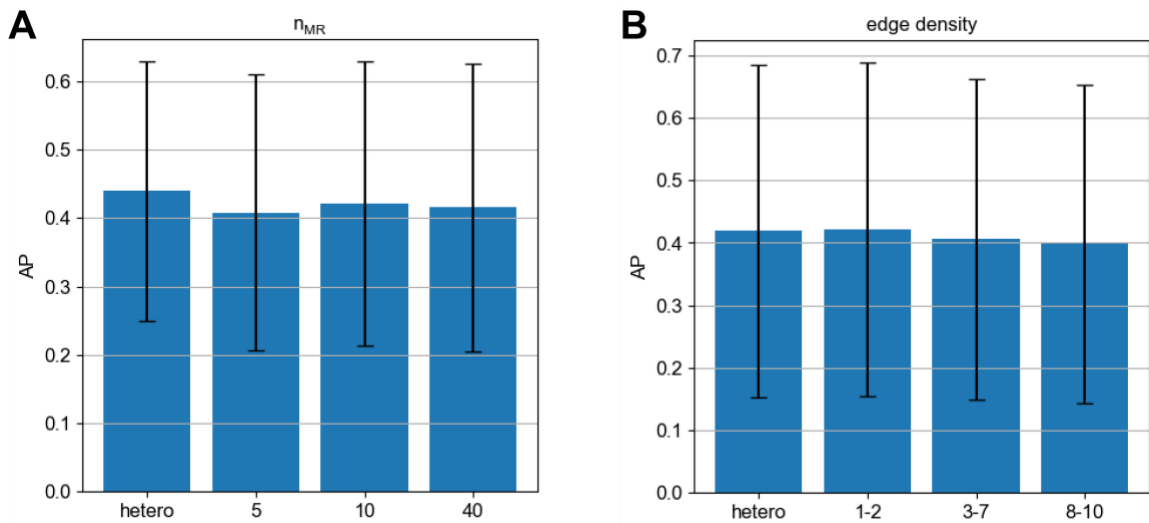
Supplementary Figure 2. Average Precision of SPREd-ML when using all but one (left) or only one (right) of the five features describing each TF-gene or TF-TF pair. AP when using all five features is shown in blue dashed line.



Supplementary Figure 3. AP scores of SPREd-ML for varying benchmark parameters. Performance (average precision) of SPREd-ML with varying edge density ($d_{TF \rightarrow G}$) of 1-2, 3-7, and 8-10 TFs per target gene (A), varying numbers of MRs (n_{MR}) (B), varying numbers of TFs (n_{TF}) (C), and varying levels of dropout added to the synthetic expression matrix (D).

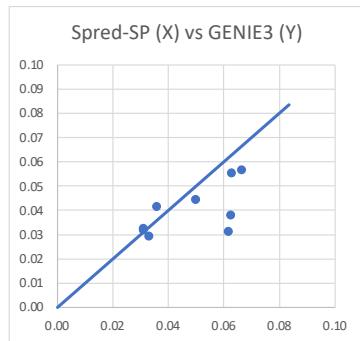
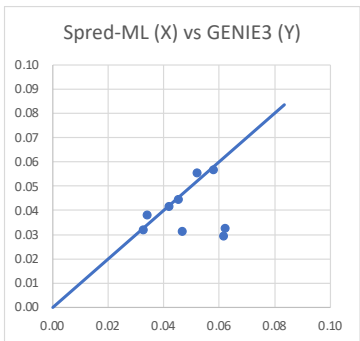
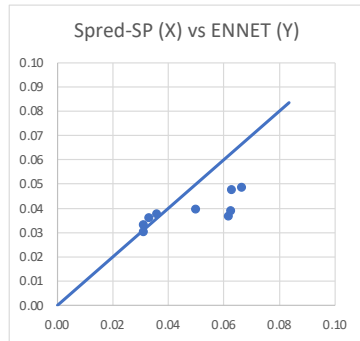
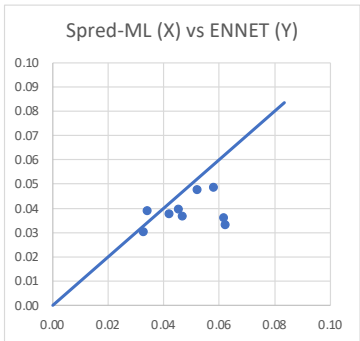
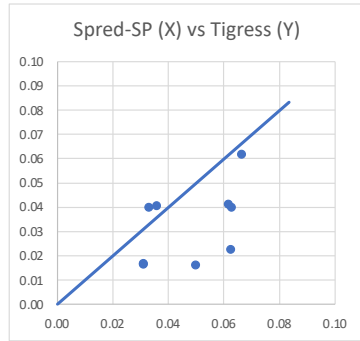
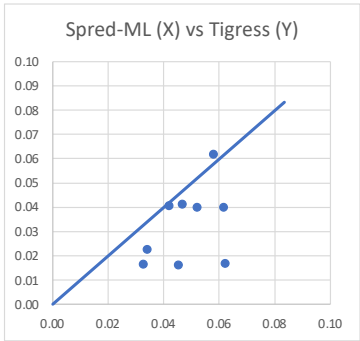
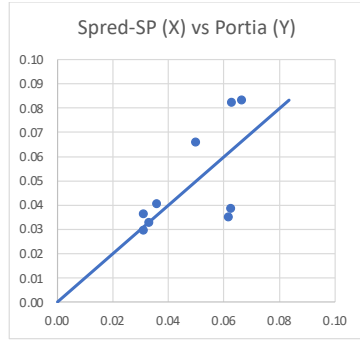
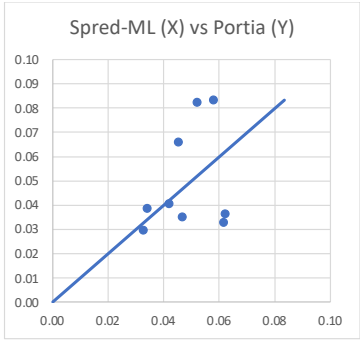


Supplementary Figure 4. SPRED performance (with default GRN parameters of $n_{MR} = 5$, $n_{TF} = 100$, $d_{MR \rightarrow TF} = 2$) at different values of edge density $d_{TF \rightarrow G} = 3, 4, \dots, 7$. Average Precision (AP) is shown in panel (A) and AUROC is shown in panel (B). The blue line in (A) shows the random expectation of AP.



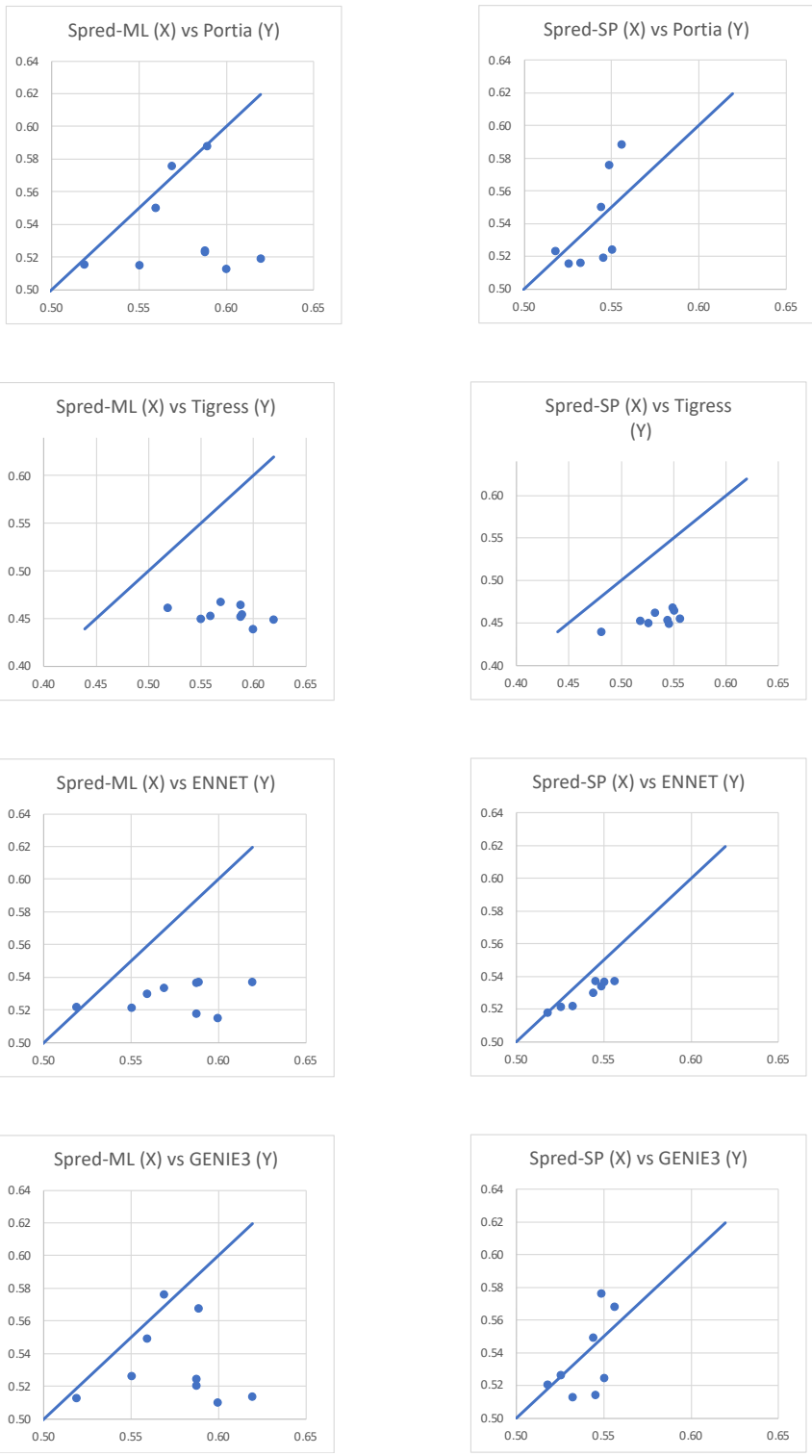
Supplementary Figure 5. Performance comparison on heterogeneous benchmarks. (A) Performance of SPRED-SP on heterogeneous test sets of different using model weights from training sets with n_{MR} set to 5, 10, and 40 or a mix thereof (“hetero”). (B) Performance of SPRED-SP on heterogeneous test sets of varying edge densities

using model weights trained on datasets with different edge densities – 1-2, 3-7, 8-10, or the entire range (“hetero”).



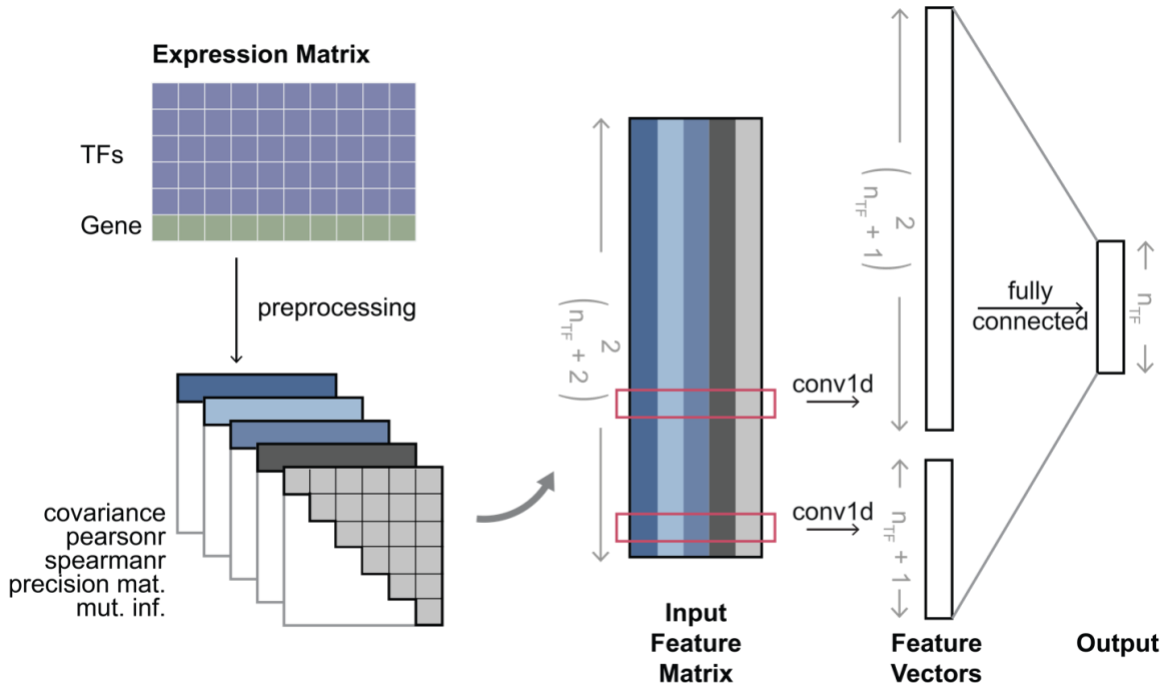
Supplementary Figure 6. Head-to-head comparison of SPRED-ML (left) or SPRED-SP (right) versus each comparator (PORTIA, TIGRESS, ENNET, GENIE3) in terms of Average Precision (AP) on each of the nine benchmarks (“tests”). In each panel,

the nine points shown represent the tests, X-axis is the AP (averaged over all genes) of a SPREd model and Y-axis is the AP of a comparator.

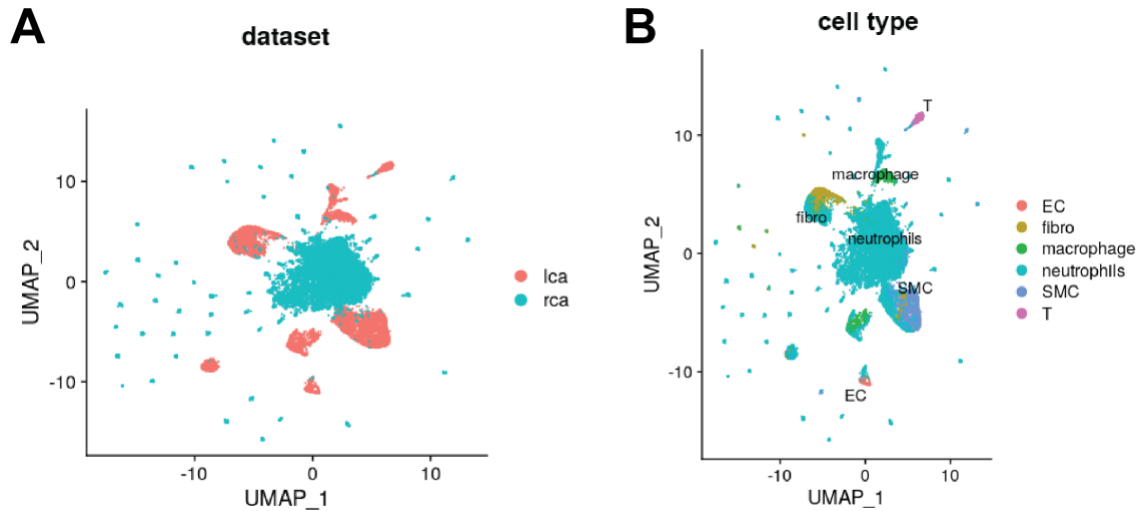


Supplementary Figure 7. Head-to-head comparison of SPRED-ML (left) or SPRED-SP (right) versus each comparator (PORTIA, TIGRESS, ENNET, GENIE3) in terms of AUROC on each of the nine benchmarks (“tests”). In each panel, the nine points

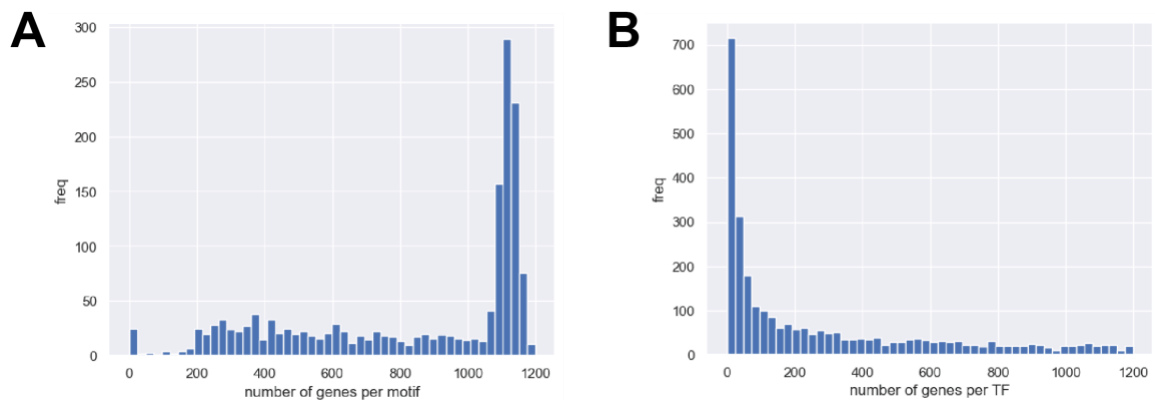
shown represent the benchmarks, X-axis is the AUROC (averaged over all genes) of a SPREd model and Y-axis is the AUROC of a comparator.



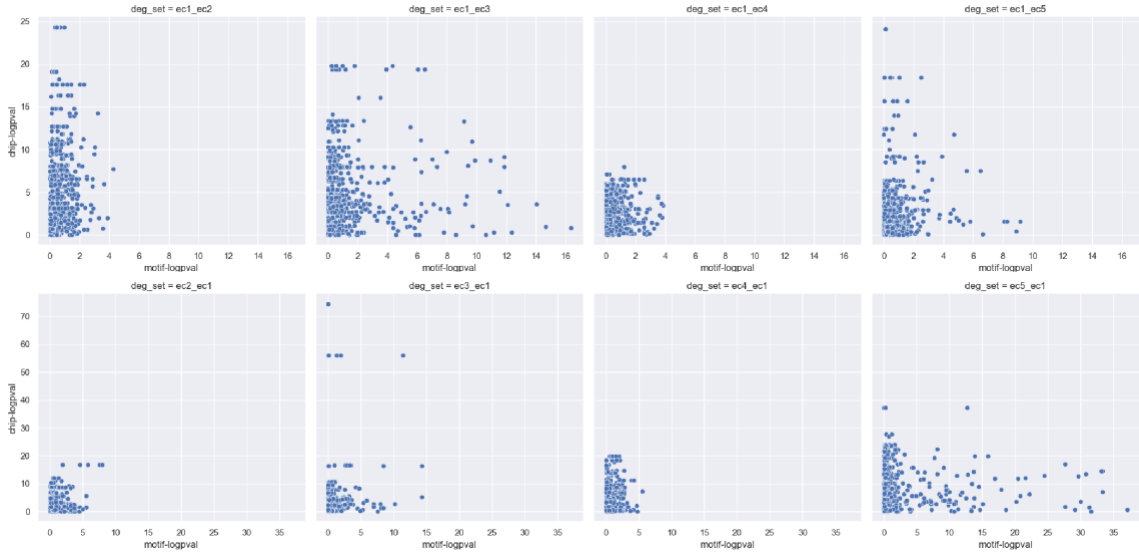
Supplementary Figure 8. Architecture of SPREd-ML neural network model. Given an expression matrix whose rows represent n_{TF} TFs and one target gene, the preprocessing step creates five features for each of the $\binom{n_{TF} + 1}{2}$ TF-TF pairs and each of the $n_{TF} + 1$ TF-target gene pairs. These features include covariance, Pearson correlation, Spearman correlation, mutual information and precision matrix entry corresponding to the TF-TF or TF-target gene pair. The five features of every gene pair then serve as the inputs of a 1D convolutional neural network (CNN). The output layer consists of n_{TF} binary labels indicating if each TF is a regulator of the target gene.



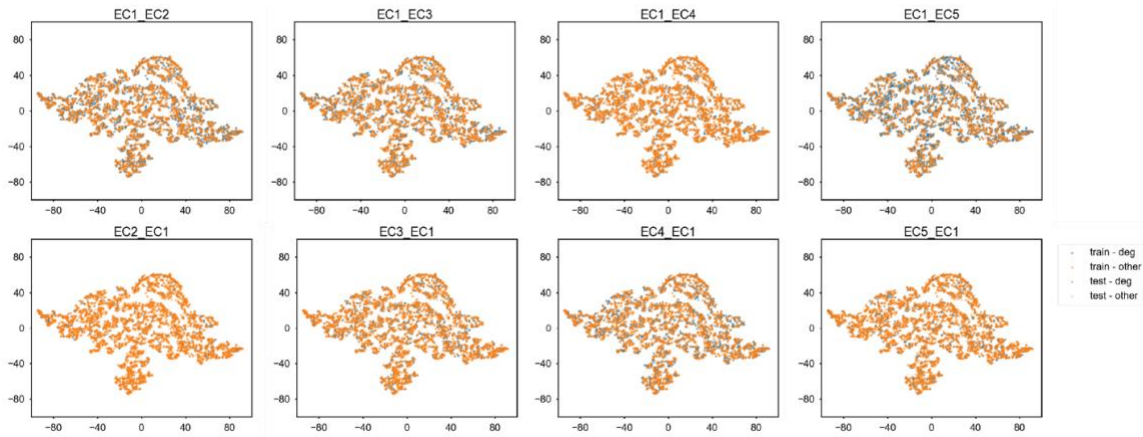
Supplementary Figure 9. UMAP plots of integrated 4 week PCL with hypercholesterolemia (D-flow_HighChol) scATAC-seq dataset including cells collected from LCA and RCA. (A) was grouped by experiment condition and showed the sources of dataset. The scattered clusters were observed in the dataset derived from RCA. (B) was grouped by cell types with labels transferred from integrated and annotated scRNA-seq data. The cell population was dominant by neutrophils, suggesting they present the major cell type in RCA when analyzed together with (A).



Supplementary Figure 10. Histogram of the number of genes per TF using motif- (A) and ChIP- (B) based methods.



Supplementary Figure 11. Comparisons of the TF-DEG association indicated by $-\log_{10}$ p-values from motif analysis (x-axis) and CHIP analysis (y-axis). The scatter plots shown here are detailed views of Figure 2D, with the data split and visualized by each DEG set.



Supplementary Figure 12. Baseline model feature visualization in tSNE plots. For validation of our proposed classifier, we did negative control and randomly shuffled the labels as the baseline model. The dataset was then partitioned into a 80% training set and a 20% testing set. the classifier was then re-trained exclusively on the training data. was trained and tested using shuffled labels of the integrated input data. The tSNE plots illustrates the hidden layer before output layer showing all data points.

APPENDIX B. SUPPLYMENTARY TABLES

Supplementary Table 1. P-values of paired Wilcoxon test (two-tailed) of difference in AP scores (A) and AUROC (B) on 5000 genes between SPREd-SP and PORTIA, ENNET, GENIE3, or TIGRESS. Green/red font indicates better/worse SPREd-SP performance.

(A)

	PORTIA	ENNET	GENIE3	TIGRESS
100	6.74E-63	6.29E-45	3.68E-106	2.42E-71
50	0.0024	4.34E-179	1.15E-245	2.74E-144
25	2.01E-07	1.92E-175	5.13E-161	1.33E-68
10	8.89E-15	0	9.90E-108	-

(B)

	PORTIA	ENNET	GENIE3	TIGRESS
100	1.09E-26	1.07E-05	4.98E-44	4.28E-43
50	0.00043	8.32E-97	3.29E-160	9.16E-123
25	0.00055	9.97E-108	1.28E-117	8.00E-65
10	0.0024	3.95E-250	1.02E-51	-

Supplementary Table 2. Summaries of different networks in the DREAM5 benchmark, including the number of TFs (n_{TF}), the number of total genes (n_{genes}), the number of conditions (n_{cond}), and the number of target genes ($n_{targets}$) for each network. Source: <https://www.synapse.org/#!/Synapse:syn2787209/wiki/70350>

	n_{TF}	n_{genes}	n_{cond}	$n_{targets}$
Network1	195	1643	805	1387
Network2	99	2810	160	367
Network3	334	4511	805	922
Network4	333	5950	536	1798

Supplementary Table 3. Performance evaluation on different networks from the DREAM5 benchmark. We compared the performance of different methods in terms of Average Precision (AP) (A) and AUROC (B). TIGRESS was not included here since it has a runtime issue and the runs failed to complete.

(A)	SPREd-SP	PORTIA	GENIE3	ENNET
Network1	0.31	0.52	0.58	0.035
Network2	0.16	0.12	0.13	0.055
Network3	0.20	0.17	0.16	0.017
Network4	0.053	0.039	0.034	0.017

(B)	SPREd-SP	PORTIA	GENIE3	ENNET
Network1	0.71	0.85	0.87	0.50
Network2	0.61	0.61	0.63	0.53
Network3	0.67	0.67	0.71	0.49
Network4	0.54	0.54	0.55	0.50

Supplementary Table 4. Run-time and scalability of SPREd-SP. (A) Run time for training the SPREd model, per epoch, for varying data sizes ($n_{TF} \times n_{genes}$). A full training typically spans 50 epochs. (B) Run time for applying a trained SPREd model to infer GRN for data sets of varying sizes.

(A) Runtime for training

$n_{TF} \times n_{genes}$	Time (s) per epoch
2500000	378.45
1500000	222.87
500000	76.741

(B) Runtime for testing

$n_{TF} \times n_{genes}$	Time (s)
500000	558.43
100000	109.45
50000	54.52

Supplementary Table 5. Comparison of AP score and AUROC between SPREd-SP CNN architecture and simple 2-layer MLP. Pair-wise Wilcoxon p-values comparing SPREd and MLP model are listed at the bottom.

	AP	AUROC
<u>SPREd</u>	0.35	0.76
MLP	0.26	0.70
p-value	0	1.23e-310

Supplementary Table 6. Mean of “fold change AP” of SPREd-ML over all genes, for each of the nine benchmarks. Fold change AP refers to the observed AP divided by the random baseline AP for a gene.

	Maclsaac2	Yeastract Type2	Yeastract Count3
NatVar	1.54	1.74	2.14
KO	1.19	2.28	1.66
Stress	2.39	1.54	1.25

Supplementary Table 7. top 2 paired sgRNA oligonucleotides targeting the promoter region.

Target Gene	sgRNA 1	sgRNA 2
CREB3L2	GTGGAGCCGTAGCGGCAAGT	GAGCAGAGCCGAGGAGCCCT
MEF2C	GTAGCAGCCCGAAGATGTCT	GATCAGGGGTACTTACATGA
RELA	GTCGTCTGTAGTGCACGCCG	GCATTTCCGCCTCTGGCGAA
NT	GTATTACTGATATTGGTGGG	

Supplementary Table 8. Primers used for target genes.

Primer Name	Nucleotide Sequence
CREB3L2 Forward	CAGAGAAGAGTGTGTCAATGGAG
CREB3L2 Reverse	CTGGTGGTAATGTGGGTGAAG
MEF2C Forward	CCAACCTTCGAGATGCCAGTCT
MEF2C Reverse	GTCGATGTGTTACACCAGGAG
RELA Forward	TGAACCGAAACTCTGGCAGCTG
RELA Reverse	CATCAGCTTGCGAAAAGGAGCC
GAPDH Forward	GTCTCCTCTGACTTCAACAGCG
GAPDH Reverse	ACCACCCTGTTGCTGTAGCCAA

Supplementary Table 9. Fused classifier evaluations on testing dataset including accuracy, precision, recall, and F1 score. (A) shows the results from our proposed model and (B) is from the baseline model that was trained and tested using randomly shuffled labels. The ratio of the number of positive labels (the sample is differentially expressed between EC1 and EC*) and negative labels are shown in the last row.

(A)

	EC1_EC2	EC1_EC3	EC1_EC4	EC1_EC5	EC2_EC1	EC3_EC1	EC4_EC1	EC5_EC1	Average
Accuracy	76.2712	76.9691	75.9721	84.2473	94.317	82.4526	79.4616	88.8335	82.3156
Precision	0.6414	0.6133	0.3613	0.8046	0.6164	0.5561	0.6406	0.6596	
Recall	0.7729	0.7709	0.6	0.9042	0.6081	0.595	0.7294	0.7209	
F1_score	0.701	0.6831	0.451	0.8515	0.6122	0.5749	0.6821	0.6889	
pos/neg	0.347	0.307	0.166	0.476	0.072	0.197	0.304	0.161	

(B)

	EC1_EC2	EC1_EC3	EC1_EC4	EC1_EC5	EC2_EC1	EC3_EC1	EC4_EC1	EC5_EC1	Average
Accuracy	33.998	45.1645	62.7119	50.349	81.1565	30.4088	34.3968	33.1007	46.4108
Precision	0.3176	0.3003	0.131	0.4868	0.0336	0.1797	0.3104	0.1631	
Recall	0.8815	0.6267	0.2375	0.855	0.0513	0.7396	0.854	0.7546	
F1_score	0.467	0.406	0.1689	0.6204	0.0406	0.2892	0.4553	0.2683	
pos/neg	0.347	0.307	0.166	0.476	0.072	0.197	0.304	0.161	

REFERENCES

- Aibar, S., Gonzalez-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J. C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., & Aerts, S. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nat Methods*, *14*(11), 1083-1086. <https://doi.org/10.1038/nmeth.4463>
- Andueza, A., Kumar, S., Kim, J., Kang, D. W., Mumme, H. L., Perez, J. I., Villa-Roel, N., & Jo, H. (2020). Endothelial Reprogramming by Disturbed Flow Revealed by Single-Cell RNA and Chromatin Accessibility Study. *Cell Rep*, *33*(11), 108491. <https://doi.org/10.1016/j.celrep.2020.108491>
- Bagnara, J. T. (1964). [Analysis of Transformations of Skin Pteridines during the Larval Life and Metamorphosis in the Triton Pleurodeles Waltlii Michah. Changes Induced by the Localized Action of Thyroxin-Cholesterol Implants]. *C R Hebd Seances Acad Sci*, *258*, 5969-5971. <https://www.ncbi.nlm.nih.gov/pubmed/14184328> (Analyse des transformations des p'teridines de la peau au cours de la vie larvaire et 'a la m'etamorphose chez le triton pleurodeles waltlii michah. Changements induits par l'action localis'ee d'implants de thyroxine-cholest'erol.)
- Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Chamberlain, A. M., Chang, A. R., Cheng, S., Das, S. R., Delling, F. N., Djousse, L., Elkind, M. S. V., Ferguson, J. F., Fornage, M., Jordan, L. C., Khan, S. S., Kissela, B. M., Knutson, K. L., . . . Stroke Statistics, S. (2019). Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation*, *139*(10), e56-e528. <https://doi.org/10.1161/CIR.0000000000000659>
- Bradner, J. E., Hnisz, D., & Young, R. A. (2017). Transcriptional Addiction in Cancer. *Cell*, *168*(4), 629-643. <https://doi.org/10.1016/j.cell.2016.12.013>
- Bravo Gonzalez-Blas, C., De Winter, S., Hulselmans, G., Hecker, N., Matetovici, I., Christiaens, V., Poovathingal, S., Wouters, J., Aibar, S., & Aerts, S. (2023). SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods*, *20*(9), 1355-1367. <https://doi.org/10.1038/s41592-023-01938-4>
- Chan, T. E., Stumpf, M. P. H., & Babbie, A. C. (2017). Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst*, *5*(3), 251-267 e253. <https://doi.org/10.1016/j.cels.2017.08.014>

- Chen, J., Cheong, C., Lan, L., Zhou, X., Liu, J., Lyu, A., Cheung, W. K., & Zhang, L. (2021). DeepDRIM: a deep neural network to reconstruct cell-type-specific gene regulatory network using single-cell RNA-seq data. *Brief Bioinform*, 22(6). <https://doi.org/10.1093/bib/bbab325>
- Chiu, J. J., & Chien, S. (2011). Effects of disturbed flow on vascular endothelium: pathophysiological basis and clinical perspectives. *Physiol Rev*, 91(1), 327-387. <https://doi.org/10.1152/physrev.00047.2009>
- Consortium, G. T. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, 45(6), 580-585. <https://doi.org/10.1038/ng.2653>
- Davidson, E. H., & Erwin, D. H. (2006). Gene regulatory networks and the evolution of animal body plans. *Science*, 311(5762), 796-800. <https://doi.org/10.1126/science.1113832>
- de Winther, M. P., Kanters, E., Kraal, G., & Hofker, M. H. (2005). Nuclear factor kappaB signaling in atherogenesis. *Arterioscler Thromb Vasc Biol*, 25(5), 904-914. <https://doi.org/10.1161/01.ATV.0000160340.72641.87>
- Dibaenia, P., & Sinha, S. (2020). SERGIO: A Single-Cell Expression Simulator Guided by Gene Regulatory Networks. *Cell Syst*, 11(3), 252-271 e211. <https://doi.org/10.1016/j.cels.2020.08.003>
- Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J., & Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol*, 34(2), 184-191. <https://doi.org/10.1038/nbt.3437>
- Duan, M., Nguyen, D. C., Joyner, C. J., Saney, C. L., Tipton, C. M., Andrews, J., Lonial, S., Kim, C., Hentenaar, I., Kusters, A., Ghosn, E., Jackson, A., Knechtle, S., Maruthamuthu, S., Chandran, S., Martin, T., Rajalingam, R., Vincenti, F., Breeden, C., . . . Lee, F. E. (2023). Understanding heterogeneity of human bone marrow plasma cell maturation and survival pathways by single-cell analyses. *Cell Rep*, 42(7), 112682. <https://doi.org/10.1016/j.celrep.2023.112682>
- Dunn, J., Qiu, H., Kim, S., Jjingo, D., Hoffman, R., Kim, C. W., Jang, I., Son, D. J., Kim, D., Pan, C., Fan, Y., Jordan, I. K., & Jo, H. (2014). Flow-dependent epigenetic DNA methylation regulates endothelial gene expression and atherosclerosis. *J Clin Invest*, 124(7), 3187-3199. <https://doi.org/10.1172/JCI74792>
- Emad, A., & Sinha, S. (2021). Inference of phenotype-relevant transcriptional regulatory networks elucidates cancer type-specific regulatory mechanisms in a pan-cancer study. *NPJ Syst Biol Appl*, 7(1), 9. <https://doi.org/10.1038/s41540-021-00169-7>

- Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., & Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun*, *10*(1), 390. <https://doi.org/10.1038/s41467-018-07931-2>
- Fernandez, D. M., Rahman, A. H., Fernandez, N. F., Chudnovskiy, A., Amir, E. D., Amadori, L., Khan, N. S., Wong, C. K., Shamailova, R., Hill, C. A., Wang, Z., Remark, R., Li, J. R., Pina, C., Faries, C., Awad, A. J., Moss, N., Bjorkegren, J. L. M., Kim-Schulze, S., . . . Giannarelli, C. (2019). Single-cell immune landscape of human atherosclerotic plaques. *Nat Med*, *25*(10), 1576-1588. <https://doi.org/10.1038/s41591-019-0590-4>
- Fleck, J. S., Jansen, S. M. J., Wollny, D., Zenk, F., Seimiya, M., Jain, A., Okamoto, R., Santel, M., He, Z., Camp, J. G., & Treutlein, B. (2023). Inferring and perturbing cell fate regulomes in human brain organoids. *Nature*, *621*(7978), 365-372. <https://doi.org/10.1038/s41586-022-05279-8>
- Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J Comput Biol*, *7*(3-4), 601-620. <https://doi.org/10.1089/106652700750050961>
- Glass, C. K., & Witztum, J. L. (2001). Atherosclerosis. the road ahead. *Cell*, *104*(4), 503-516. [https://doi.org/10.1016/s0092-8674\(01\)00238-0](https://doi.org/10.1016/s0092-8674(01)00238-0)
- Gold, E. S., Ramsey, S. A., Sartain, M. J., Selinummi, J., Podolsky, I., Rodriguez, D. J., Moritz, R. L., & Aderem, A. (2012). ATF3 protects against atherosclerosis by suppressing 25-hydroxycholesterol-induced lipid body formation. *J Exp Med*, *209*(4), 807-817. <https://doi.org/10.1084/jem.20111202>
- Guo, S., Jiang, Q., Chen, L., & Guo, D. (2016). Gene regulatory network inference using PLS-based methods. *BMC Bioinformatics*, *17*(1), 545. <https://doi.org/10.1186/s12859-016-1398-6>
- Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., & Satija, R. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol*, *42*(2), 293-304. <https://doi.org/10.1038/s41587-023-01767-y>
- Haury, A. C., Mordelet, F., Vera-Licona, P., & Vert, J. P. (2012). TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst Biol*, *6*, 145. <https://doi.org/10.1186/1752-0509-6-145>
- Herrington, W., Lacey, B., Sherliker, P., Armitage, J., & Lewington, S. (2016). Epidemiology of Atherosclerosis and the Potential to Reduce the Global Burden of Atherothrombotic Disease. *Circ Res*, *118*(4), 535-546. <https://doi.org/10.1161/CIRCRESAHA.115.307611>
- Horlbeck, M. A., Gilbert, L. A., Villalta, J. E., Adamson, B., Pak, R. A., Chen, Y., Fields, A. P., Park, C. Y., Corn, J. E., Kampmann, M., & Weissman, J. S. (2016). Compact

- and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife*, 5. <https://doi.org/10.7554/eLife.19760>
- Huang, X., Brown, C., Ni, W., Maynard, E., Rigby, A. C., & Oettgen, P. (2006). Critical role for the Ets transcription factor ELF-1 in the development of tumor angiogenesis. *Blood*, 107(8), 3153-3160. <https://doi.org/10.1182/blood-2005-08-3206>
- Huynh-Thu, V. A., & Geurts, P. (2018). dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. *Sci Rep*, 8(1), 3384. <https://doi.org/10.1038/s41598-018-21715-0>
- Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, 5(9). <https://doi.org/10.1371/journal.pone.0012776>
- Jebari-Benslaiman, S., Galicia-Garcia, U., Larrea-Sebal, A., Olaetxea, J. R., Alloza, I., Vandenbroeck, K., Benito-Vicente, A., & Martin, C. (2022). Pathophysiology of Atherosclerosis. *Int J Mol Sci*, 23(6). <https://doi.org/10.3390/ijms23063346>
- Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830), 1497-1502. <https://doi.org/10.1126/science.1141319>
- Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., & Morris, S. A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949), 742-751. <https://doi.org/10.1038/s41586-022-05688-9>
- Kardassis, D., Vindis, C., Stancu, C. S., Toma, L., Gafencu, A. V., Georgescu, A., Alexandru-Moise, N., Molica, F., Kwak, B. R., Burlacu, A., Hall, I. F., Butoi, E., Magni, P., Wu, J., Novella, S., Gamon, L. F., Davies, M. J., Caporali, A., de la Cuesta, F., & Mitic, T. (2025). Unravelling molecular mechanisms in atherosclerosis using cellular models and omics technologies. *Vascul Pharmacol*, 158, 107452. <https://doi.org/10.1016/j.vph.2024.107452>
- Kazemian, M., Blatti, C., Richards, A., McCutchan, M., Wakabayashi-Ito, N., Hammonds, A. S., Celniker, S. E., Kumar, S., Wolfe, S. A., Brodsky, M. H., & Sinha, S. (2010). Quantitative analysis of the *Drosophila* segmentation regulatory network using pattern generating potentials. *PLoS Biol*, 8(8). <https://doi.org/10.1371/journal.pbio.1000456>
- Kharchenko, P. V., Silberstein, L., & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat Methods*, 11(7), 740-742. <https://doi.org/10.1038/nmeth.2967>

- Kim, D., Tran, A., Kim, H. J., Lin, Y., Yang, J. Y. H., & Yang, P. (2023). Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data. *NPJ Syst Biol Appl*, 9(1), 51. <https://doi.org/10.1038/s41540-023-00312-6>
- Kolmykov, S., Yevshin, I., Kulyashov, M., Sharipov, R., Kondrakhin, Y., Makeev, V. J., Kulakovskiy, I. V., Kel, A., & Kolpakov, F. (2021). GTRD: an integrated view of transcription regulation. *Nucleic Acids Res*, 49(D1), D104-D111. <https://doi.org/10.1093/nar/gkaa1057>
- Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Mol Cell*, 58(4), 610-620. <https://doi.org/10.1016/j.molcel.2015.04.005>
- Kumar, S., Kim, C. W., Simmons, R. D., & Jo, H. (2014). Role of flow-sensitive microRNAs in endothelial dysfunction and atherosclerosis: mechanosensitive athero-miRs. *Arterioscler Thromb Vasc Biol*, 34(10), 2206-2216. <https://doi.org/10.1161/ATVBAHA.114.303425>
- Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 559. <https://doi.org/10.1186/1471-2105-9-559>
- Lareau, C. A., Duarte, F. M., Chew, J. G., Kartha, V. K., Burkett, Z. D., Kohlway, A. S., Pokholok, D., Aryee, M. J., Steemers, F. J., Lebofsky, R., & Buenrostro, J. D. (2019). Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat Biotechnol*, 37(8), 916-924. <https://doi.org/10.1038/s41587-019-0147-6>
- Lavillegrand, J. R., Al-Rifai, R., Thietart, S., Guyon, T., Vandestienne, M., Cohen, R., Duval, V., Zhong, X., Yen, D., Ozturk, M., Negishi, Y., Konkeli, J., Pinteaux, E., Lenoir, O., Vilar, J., Laurans, L., Esposito, B., Bredon, M., Sokol, H., . . . Ait-Oufella, H. (2024). Alternating high-fat diet enhances atherosclerosis by neutrophil reprogramming. *Nature*, 634(8033), 447-456. <https://doi.org/10.1038/s41586-024-07693-6>
- Lee, W. P., & Tzou, W. S. (2009). Computational methods for discovering gene networks from expression data. *Brief Bioinform*, 10(4), 408-423. <https://doi.org/10.1093/bib/bbp028>
- Levine, M., & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424(6945), 147-151. <https://doi.org/10.1038/nature01763>
- Li, H., Zhang, Z., Squires, M., Chen, X., & Zhang, X. (2023). scMultiSim: simulation of single cell multi-omics and spatial data guided by gene regulatory networks and cell-cell interactions. *Res Sq*. <https://doi.org/10.21203/rs.3.rs-3301625/v1>

- Li, W. V., & Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun*, 9(1), 997. <https://doi.org/10.1038/s41467-018-03405-7>
- Li, X., Ma, S., Liu, J., Tang, J., & Guo, F. (2022). Inferring gene regulatory network via fusing gene expression image and RNA-seq data. *Bioinformatics*, 38(6), 1716-1723. <https://doi.org/10.1093/bioinformatics/btac008>
- Liang, J., & Han, J. (2012). Stochastic Boolean networks: an efficient approach to modeling gene regulatory networks. *BMC Syst Biol*, 6, 113. <https://doi.org/10.1186/1752-0509-6-113>
- Libby, P. (2021). The changing landscape of atherosclerosis. *Nature*, 592(7855), 524-533. <https://doi.org/10.1038/s41586-021-03392-8>
- Libby, P., Ridker, P. M., Hansson, G. K., & Leducq Transatlantic Network on, A. (2009). Inflammation in atherosclerosis: from pathophysiology to practice. *J Am Coll Cardiol*, 54(23), 2129-2138. <https://doi.org/10.1016/j.jacc.2009.09.009>
- Lim, N., Senbabaoglu, Y., Michailidis, G., & d'Alche-Buc, F. (2013). OKVAR-Boost: a novel boosting algorithm to infer nonlinear dynamics and interactions in gene regulatory networks. *Bioinformatics*, 29(11), 1416-1423. <https://doi.org/10.1093/bioinformatics/btt167>
- Lin, Z., & Ou-Yang, L. (2023). Inferring gene regulatory networks from single-cell gene expression data via deep multi-view contrastive learning. *Brief Bioinform*, 24(1). <https://doi.org/10.1093/bib/bbac586>
- Luo, C., Ruan, Y., Sun, P., Wang, H., Yang, W., Gong, Y., & Wang, D. (2022). The Role of Transcription Factors in Coronary Artery Disease and Myocardial Infarction. *Front Biosci (Landmark Ed)*, 27(12), 329. <https://doi.org/10.31083/j.fbl2712329>
- MacIsaac, K. D., Wang, T., Gordon, D. B., Gifford, D. K., Stormo, G. D., & Fraenkel, E. (2006). An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, 7, 113. <https://doi.org/10.1186/1471-2105-7-113>
- Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Consortium, D., Kellis, M., Collins, J. J., & Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nat Methods*, 9(8), 796-804. <https://doi.org/10.1038/nmeth.2016>
- Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., & Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Methods*, 13(4), 366-370. <https://doi.org/10.1038/nmeth.3799>
- Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene

- regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1(Suppl 1), S7. <https://doi.org/10.1186/1471-2105-7-S1-S7>
- Markowetz, F., & Spang, R. (2007). Inferring cellular networks--a review. *BMC Bioinformatics*, 8 Suppl 6(Suppl 6), S5. <https://doi.org/10.1186/1471-2105-8-S6-S5>
- Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S. H., Ko, S. B. H., Gouda, N., Hayashi, T., & Nikaido, I. (2017). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*, 33(15), 2314-2321. <https://doi.org/10.1093/bioinformatics/btx194>
- McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Moerman, T., Aibar Santos, S., Bravo Gonzalez-Blas, C., Simm, J., Moreau, Y., Aerts, J., & Aerts, S. (2019). GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12), 2159-2161. <https://doi.org/10.1093/bioinformatics/bty916>
- Nam, D., Ni, C. W., Rezvan, A., Suo, J., Budzyn, K., Llanos, A., Harrison, D., Giddens, D., & Jo, H. (2009). Partial carotid ligation is a model of acutely induced disturbed flow, leading to rapid endothelial dysfunction and atherosclerosis. *Am J Physiol Heart Circ Physiol*, 297(4), H1535-1543. <https://doi.org/10.1152/ajpheart.00510.2009>
- Ord, T., Lonngberg, T., Nurminen, V., Ravindran, A., Niskanen, H., Kiema, M., Ounap, K., Maria, M., Moreau, P. R., Mishra, P. P., Palani, S., Virta, J., Liljenback, H., Aavik, E., Roivainen, A., Yla-Herttuala, S., Laakkonen, J. P., Lehtimaki, T., & Kaikkonen, M. U. (2023). Dissecting the polygenic basis of atherosclerosis via disease-associated cell state signatures. *Am J Hum Genet*, 110(5), 722-740. <https://doi.org/10.1016/j.ajhg.2023.03.013>
- Park, C., Baek, K. I., Hung, R. C., Choi, L., Jeong, K., Kim, P., Jahng, A. K., Kim, J. H., Meselhe, M., Kannan, A., Chou, C. L., Kang, D. W., Song, E. J., Kim, Y., Bowman-Kirigin, J. A., Clark, M. D., van der Laan, S. W., Pasterkamp, G., Villa-Roel, N., . . . Jo, H. (2025). Disturbed Flow Induces Reprogramming of Endothelial Cells to Immune-like and Foam Cells under Hypercholesterolemia during Atherogenesis. *bioRxiv*. <https://doi.org/10.1101/2025.03.06.641843>
- Passemiers, A., Moreau, Y., & Raimondi, D. (2022). Fast and accurate inference of gene regulatory networks through robust precision matrix estimation. *Bioinformatics*, 38(10), 2802-2809. <https://doi.org/10.1093/bioinformatics/btac178>
- Peng, J., Le, C. Y., Xia, B., Wang, J. W., Liu, J. J., Li, Z., Zhang, Q. J., Zhang, Q., Wang, J., & Wan, C. W. (2021). Research on the correlation between activating transcription factor 3 expression in the human coronary artery and atherosclerotic

- plaque stability. *BMC Cardiovasc Disord*, 21(1), 356. <https://doi.org/10.1186/s12872-021-02161-9>
- Perrin, B. E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., & d'Alche-Buc, F. (2003). Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, 19 Suppl 2, ii138-148. <https://doi.org/10.1093/bioinformatics/btg1071>
- Peter, I. S., & Davidson, E. H. (2011). A gene regulatory network controlling the embryonic specification of endoderm. *Nature*, 474(7353), 635-639. <https://doi.org/10.1038/nature10100>
- Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J., & Trapnell, C. (2018). Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell*, 71(5), 858-871 e858. <https://doi.org/10.1016/j.molcel.2018.06.044>
- Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods*, 17(2), 147-154. <https://doi.org/10.1038/s41592-019-0690-6>
- Qureshi, I. A., & Mehler, M. F. (2012). Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. *Nat Rev Neurosci*, 13(8), 528-541. <https://doi.org/10.1038/nrn3234>
- Romanoski, C. E., Lee, S., Kim, M. J., Ingram-Drake, L., Plaisier, C. L., Yordanova, R., Tilford, C., Guan, B., He, A., Gargalovic, P. S., Kirchgessner, T. G., Berliner, J. A., & Lusk, A. J. (2010). Systems genetics analysis of gene-by-environment interactions in human cells. *Am J Hum Genet*, 86(3), 399-410. <https://doi.org/10.1016/j.ajhg.2010.02.002>
- Ruysinck, J., Huynh-Thu, V. A., Geurts, P., Dhaene, T., Demeester, P., & Saeys, Y. (2014). NIMEFI: gene regulatory network inference using multiple ensemble feature importance algorithms. *PLoS One*, 9(3), e92709. <https://doi.org/10.1371/journal.pone.0092709>
- Sanson, K. R., Hanna, R. E., Hegde, M., Donovan, K. F., Strand, C., Sullender, M. E., Vaimberg, E. W., Goodale, A., Root, D. E., Piccioni, F., & Doench, J. G. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nat Commun*, 9(1), 5416. <https://doi.org/10.1038/s41467-018-07901-8>
- Saul, M. C., Seward, C. H., Troy, J. M., Zhang, H., Sloofman, L. G., Lu, X., Weisner, P. A., Caetano-Anolles, D., Sun, H., Zhao, S. D., Chandrasekaran, S., Sinha, S., & Stubbs, L. (2017). Transcriptional regulatory dynamics drive coordinated metabolic and neural response to social challenge in mice. *Genome Res*, 27(6), 959-972. <https://doi.org/10.1101/gr.214221.116>

- Schaffter, T., Marbach, D., & Floreano, D. (2011). GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16), 2263-2270. <https://doi.org/10.1093/bioinformatics/btr373>
- Schnitzler, G. R., Kang, H., Fang, S., Angom, R. S., Lee-Kim, V. S., Ma, X. R., Zhou, R., Zeng, T., Guo, K., Taylor, M. S., Vellarikkal, S. K., Barry, A. E., Sias-Garcia, O., Bloemendal, A., Munson, G., Guckelberger, P., Nguyen, T. H., Bergman, D. T., Hinshaw, S., . . . Engreitz, J. M. (2024). Convergence of coronary artery disease genes onto endothelial cell programs. *Nature*, 626(8000), 799-807. <https://doi.org/10.1038/s41586-024-07022-x>
- Shmulevich, I., Dougherty, E. R., Kim, S., & Zhang, W. (2002). Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2), 261-274. <https://doi.org/10.1093/bioinformatics/18.2.261>
- Shrivastava, H., Zhang, X., Song, L., & Aluru, S. (2022). GRNUlar: A Deep Learning Framework for Recovering Single-Cell Gene Regulatory Networks. *J Comput Biol*, 29(1), 27-44. <https://doi.org/10.1089/cmb.2021.0437>
- Siahpirani, A. F., & Roy, S. (2017). A prior-based integrative framework for functional transcriptional regulatory network inference. *Nucleic Acids Res*, 45(4), e21. <https://doi.org/10.1093/nar/gkw963>
- Siletti, K., Hodge, R., Mossi Albiach, A., Lee, K. W., Ding, S. L., Hu, L., Lonnerberg, P., Bakken, T., Casper, T., Clark, M., Dee, N., Gloe, J., Hirschstein, D., Shapovalova, N. V., Keene, C. D., Nyhus, J., Tung, H., Yanny, A. M., Arenas, E., . . . Linnarsson, S. (2023). Transcriptomic diversity of cell types across the adult human brain. *Science*, 382(6667), eadd7046. <https://doi.org/10.1126/science.add7046>
- Sinha, S., Jones, B. M., Traniello, I. M., Bukhari, S. A., Halfon, M. S., Hofmann, H. A., Huang, S., Katz, P. S., Keagy, J., Lynch, V. J., Sokolowski, M. B., Stubbs, L. J., Tabe-Bordbar, S., Wolfner, M. F., & Robinson, G. E. (2020). Behavior-related gene regulatory networks: A new level of organization in the brain. *Proc Natl Acad Sci USA*, 117(38), 23270-23279. <https://doi.org/10.1073/pnas.1921625117>
- Skok Gibbs, C., Jackson, C. A., Saldi, G. A., Tjarnberg, A., Shah, A., Watters, A., De Veaux, N., Tchourine, K., Yi, R., Hamamsy, T., Castro, D. M., Carriero, N., Gorissen, B. L., Gresham, D., Miraldi, E. R., & Bonneau, R. (2022). High-performance single-cell gene regulatory network inference at scale: the Inferelator 3.0. *Bioinformatics*, 38(9), 2519-2528. <https://doi.org/10.1093/bioinformatics/btac117>
- Slawek, J., & Arodz, T. (2013). ENNET: inferring large gene regulatory networks from expression data using gradient boosting. *BMC Syst Biol*, 7, 106. <https://doi.org/10.1186/1752-0509-7-106>
- Son, D. J., Kumar, S., Takabe, W., Kim, C. W., Ni, C. W., Alberts-Grill, N., Jang, I. H., Kim, S., Kim, W., Won Kang, S., Baker, A. H., Woong Seo, J., Ferrara, K. W., &

- Jo, H. (2013). The atypical mechanosensitive microRNA-712 derived from pre-ribosomal RNA induces endothelial inflammation and atherosclerosis. *Nat Commun*, 4, 3000. <https://doi.org/10.1038/ncomms4000>
- Spitz, F., & Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*, 13(9), 613-626. <https://doi.org/10.1038/nrg3207>
- Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643), 249-255. <https://doi.org/10.1126/science.1087447>
- Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, 177(7), 1888-1902 e1821. <https://doi.org/10.1016/j.cell.2019.05.031>
- Sundararajan, M., Taly, A., & Yan, Q. Q. (2017). Axiomatic Attribution for Deep Networks. *International Conference on Machine Learning, Vol 70*, 70. <Go to ISI>://WOS:000683309503043
- Tabas, I., Garcia-Cardena, G., & Owens, G. K. (2015). Recent insights into the cellular biology of atherosclerosis. *J Cell Biol*, 209(1), 13-22. <https://doi.org/10.1083/jcb.201412052>
- Tamargo, I. A., Baek, K. I., Kim, Y., Park, C., & Jo, H. (2023). Flow-induced reprogramming of endothelial cells in atherosclerosis. *Nat Rev Cardiol*, 20(11), 738-753. <https://doi.org/10.1038/s41569-023-00883-1>
- Tamargo, I. A., Baek, K. I., Xu, C., Kang, D. W., Kim, Y., Andueza, A., Williams, D., Demos, C., Villa-Roel, N., Kumar, S., Park, C., Choi, R., Johnson, J., Chang, S., Kim, P., Tan, S., Jeong, K., Tsuji, S., & Jo, H. (2024). HEG1 Protects Against Atherosclerosis by Regulating Stable Flow-Induced KLF2/4 Expression in Endothelial Cells. *Circulation*, 149(15), 1183-1201. <https://doi.org/10.1161/CIRCULATIONAHA.123.064735>
- Tardajos Ayllon, B., Bowden, N., Souilhol, C., Darwish, H., Tian, S., Duckworth, C., Pritchard, D. M., Xu, S., Sayers, J., Francis, S., Serbanovic-Canic, J., Oakley, F., & Evans, P. C. (2025). Endothelial c-REL orchestrates atherosclerosis at regions of disturbed flow through crosstalk with TXNIP-p38 and non-canonical NF-kappaB pathways. *Cardiovasc Res*, 121(5), 748-759. <https://doi.org/10.1093/cvr/cvaf024>
- Tetreault, M. P., Yang, Y., & Katz, J. P. (2013). Kruppel-like factors in cancer. *Nat Rev Cancer*, 13(10), 701-713. <https://doi.org/10.1038/nrc3582>
- Traniello, I. M., Bukhari, S. A., Dibaeinia, P., Serrano, G., Avalos, A., Ahmed, A. C., Sankey, A. L., Hernaez, M., Sinha, S., Zhao, S. D., Catchen, J., & Robinson, G. E.

- (2023). Single-cell dissection of aggression in honeybee colonies. *Nat Ecol Evol*. <https://doi.org/10.1038/s41559-023-02090-0>
- van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., Bieri, B., Mazutis, L., Wolf, G., Krishnaswamy, S., & Pe'er, D. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, *174*(3), 716-729 e727. <https://doi.org/10.1016/j.cell.2018.05.061>
- van Heeringen, S. J., & Veenstra, G. J. (2011). GimmeMotifs: a de novo motif prediction pipeline for CHIP-sequencing experiments. *Bioinformatics*, *27*(2), 270-271. <https://doi.org/10.1093/bioinformatics/btq636>
- van Thienen, J. V., Fledderus, J. O., Dekker, R. J., Rohlena, J., van Ijzendoorn, G. A., Kootstra, N. A., Pannekoek, H., & Horrevoets, A. J. (2006). Shear stress sustains atheroprotective endothelial KLF2 expression more potently than statins through mRNA stabilization. *Cardiovasc Res*, *72*(2), 231-240. <https://doi.org/10.1016/j.cardiores.2006.07.008>
- Wagner, A., Regev, A., & Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol*, *34*(11), 1145-1160. <https://doi.org/10.1038/nbt.3711>
- Wang, J., Ma, A., Ma, Q., Xu, D., & Joshi, T. (2020). Inductive inference of gene regulatory network using supervised and semi-supervised graph neural networks. *Comput Struct Biotechnol J*, *18*, 3335-3343. <https://doi.org/10.1016/j.csbj.2020.10.022>
- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, *10*(1), 57-63. <https://doi.org/10.1038/nrg2484>
- Weidemuller, P., Kholmatov, M., Petsalaki, E., & Zaugg, J. B. (2021). Transcription factors: Bridge between cell signaling and gene regulation. *Proteomics*, *21*(23-24), e2000034. <https://doi.org/10.1002/pmic.202000034>
- Williams, J. W., Winkels, H., Durant, C. P., Zaitsev, K., Ghosheh, Y., & Ley, K. (2020). Single Cell RNA Sequencing in Atherosclerosis Research. *Circ Res*, *126*(9), 1112-1126. <https://doi.org/10.1161/CIRCRESAHA.119.315940>
- Wirka, R. C., Wagh, D., Paik, D. T., Pjanic, M., Nguyen, T., Miller, C. L., Kundu, R., Nagao, M., Coller, J., Koyano, T. K., Fong, R., Woo, Y. J., Liu, B., Montgomery, S. B., Wu, J. C., Zhu, K., Chang, R., Alamprese, M., Tallquist, M. D., . . . Quertermous, T. (2019). Atheroprotective roles of smooth muscle cell phenotypic modulation and the TCF21 disease gene as revealed by single-cell analysis. *Nat Med*, *25*(8), 1280-1289. <https://doi.org/10.1038/s41591-019-0512-5>

- Wu, Z., & Sinha, S. (2024). SPREd: a simulation-supervised neural network tool for gene regulatory network reconstruction. *Bioinform Adv*, 4(1), vbae011. <https://doi.org/10.1093/bioadv/vbae011>
- Xu, D., Shi, Y., Tsang, I. W., Ong, Y. S., Gong, C., & Shen, X. (2020). Survey on Multi-Output Learning. *IEEE Trans Neural Netw Learn Syst*, 31(7), 2409-2429. <https://doi.org/10.1109/TNNLS.2019.2945133>
- Xu, H., Xiang, Q. Y., Li, S., & Liu, Y. S. (2023). High serum Bhlhe40 levels are associated with subclinical atherosclerosis in patients with type 2 diabetes mellitus: A cross-sectional study. *Diab Vasc Dis Res*, 20(2), 14791641231169246. <https://doi.org/10.1177/14791641231169246>
- Yamada, Y., Nishida, T., Horibe, H., Oguri, M., Kato, K., & Sawabe, M. (2014). Identification of hypo- and hypermethylated genes related to atherosclerosis by a genome-wide analysis of DNA methylation. *Int J Mol Med*, 33(5), 1355-1363. <https://doi.org/10.3892/ijmm.2014.1692>
- Yip, K. Y., Alexander, R. P., Yan, K. K., & Gerstein, M. (2010). Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PLoS One*, 5(1), e8121. <https://doi.org/10.1371/journal.pone.0008121>
- Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., & Jarvis, E. D. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18), 3594-3603. <https://doi.org/10.1093/bioinformatics/bth448>
- Yuan, Y., & Bar-Joseph, Z. (2019). Deep learning for inferring gene relationships from single-cell expression data. *Proc Natl Acad Sci U S A*, 116(52), 27151-27158. <https://doi.org/10.1073/pnas.1911536116>
- Zheng, L., Liu, Z., Yang, Y., & Shen, H. B. (2022). Accurate inference of gene regulatory interactions from spatial gene expression with deep contrastive learning. *Bioinformatics*, 38(3), 746-753. <https://doi.org/10.1093/bioinformatics/btab718>