Biological-Network Guided Machine Learning for Understanding Gene Regulation in Human Brains and Disease Phenotypes from Multi-omics Data

By: Saniya Khullar

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

(Biomedical Data Science)

UNIVERSITY OF WISCONSIN-MADISON

2024

Date of final oral examination: August 7, 2024

This dissertation is approved by the following members of the Final Oral Committee:

Daifeng Wang, Associate Professor, Biostatistics and Medical Informatics, Computer Sciences (affiliate), Director of Data Science Core and Investigator at the Waisman Center

Mark Craven, Professor, Biostatistics and Medical Informatics, Computer Sciences, Head of the Computation and Informatics in Biology and Medicine (CIBM) program

Qiongshi Lu, Associate Professor, Affiliate Faculty, Statistics; Computer, Data & Information Sciences; College of Letters and Science

John Svaren, Distinguished Professor, Comparative Biosciences, Interim Associate Vice Chancellor for Research in the Biological Science

§ Abstract

Biological-Network Guided Machine Learning for Understanding Gene Regulation in Human Brains and Disease Phenotypes from Multi-omics Data

Saniya Khullar

Under the supervision of Professor Daifeng Wang at the University of Wisconsin-Madison

Understanding the genetic basis of nervous system diseases, such as Alzheimer's disease (AD), is crucial due to their profound impact on global health and individual quality of life. These diseases present a spectrum of clinical manifestations, with diverse phenotypes that affect their onset, progression, and severity. Genome-Wide Association Studies (GWAS) have shed light on numerous genetic variants, including Single Nucleotide Polymorphisms (SNPs), linked to AD and other brain-related disorders. However, a challenge remains to unravel molecular and cellular mechanisms by which these variants contribute to disease pathology.

A notable difficulty is that many SNPs have only modest effect sizes but can cumulatively influence the risk of brain-related disorders. Moreover, ~90% of SNPs are in non-protein-coding regions of DNA and their functional impact is difficult to decipher. Thus, it is important to interpret the role of these previously orphaned variants in the context of disease manifestation. The integration of multi-modal patient data has significantly enhanced our understanding of the molecular and cellular dynamics that underpin disease progression and phenotypic variation. Despite this progress, and the advances in computational methods integrating multi-omics data, a cohesive narrative that connects genotype and phenotype through gene regulatory mechanisms, like gene regulatory networks (GRNs), is still missing; GRNs link Transcription Factor (TF) proteins to the target genes (TGs) they help regulate. Indeed, while GWAS and expression Quantitative Trait Loci (eQTL) studies have started to map the relationships between SNPs and disease risk and SNPs and TG expression changes, respectively, the disruption of GRNs and their impact on disease TGs is not fully understood.

To help address these gaps, I introduce SNPheno, my computational pipeline developed to link disease-associated SNPs with GRNs, and TGs involved in complex phenotypes. SNPheno helps elucidate the impact of SNPs on transcriptional dysregulation by investigating how they can alter TF binding sites and, subsequently, TF regulation of TGs. These disruptions are correlated with disease phenotypes and aberrant biological processes and pathways, particularly in brain-related diseases. I apply SNPheno to

uncover genetic mechanisms associated with AD and severe Covid-19, examining 3 brain regions with known dysregulation in AD, using bulk-level data that reflects the collective patterns of diverse cell types.

The complexity of gene regulation within specific cell types necessitates a deep dive into TF coordination at the cellular level. Existing tools often overlook crucial, widespread TF-TF protein-protein interactions (PPIs) that underpin gene regulation in a cell-type-specific context. To bridge this gap, I introduce NetREm, a novel approach utilizing network regression embeddings to reveal the intricate network of TF coordination and regulatory modules regulating gene expression. NetREm integrates prior knowledge like direct and/or indirect PPIs among TFs, providing a more detailed representation of cell-type-specific GRNs and TF-TF PPIs. We benchmark NetREm's performance in various human and mouse cell types and apply it to construct both GRNs and TF-TF coordination networks in myelinating and non-myelinating human Schwann cells, as well as in eight glial/neuronal cell types in AD and control states. Top findings are validated using functional genomic data from humans, rats, and mice, including: eQTLs, GWAS, CUT&RUN (Cleavage Under Targets and Release Using Nuclease) sequencing, Chromatin immunoprecipitation sequencing (ChIP-seq) data, knockout studies.

Overall, my interpretable machine learning-based pipelines construct genotype-to-phenotype networks that identify potential candidate biomarkers for various health outcomes. This dissertation illuminates how non-coding SNPs may influence complex biological mechanisms such as TF-TF coordination, and how these mechanisms translate into changes in gene expression at the tissue or cell-type level, contributing to diverse disease phenotypes. Further, I explore the evolution of these regulatory and coordination mechanisms during disease progression and their variability across brain regions and cell types. I annotate existing TF-TF PPIs at the cell-type level and flag novel PPIs for follow-ups. By presenting SNPheno and NetREm, I offer two novel computational approaches for predicting and integrating GRN and TF-TF coordination networks from multi-omics data. These methods deepen our understanding of brain-related diseases and beyond, contribute to on-going advancements in precision medicine, and also aid in identifying network-based biomarkers for disease phenotypes.

Acknowledgements

I am deeply grateful to my advisor, Professor Daifeng Wang, for his unwavering support, invaluable mentorship, and guidance throughout my Ph.D. journey. His dedication and commitment to fostering a nurturing environment have been instrumental in my growth. Dr. Wang recognized my enthusiasm for neurodegenerative disease research and helped me refine my focus while expanding my understanding of molecular biology, bioinformatics, and neuroscience. Under his mentorship, I learned to approach research questions and challenges both systematically and creatively, breaking them down into manageable steps. Professor Wang's loyalty, humility, patience, empathy, encouraging spirit, and approachability have made him not just an advisor but a true mentor and role model. It has been an honor to work under his guidance.

I also wish to express my sincere gratitude to the other members of my Ph.D. committee for their invaluable contributions. Their diverse perspectives and sage advice have greatly enriched my research and helped significantly enhance this dissertation. A special thanks to Distinguished Professor John Svaren for being my mini-advisor within the University of Wisconsin – Madison's Computation and Informatics in Biology and Medicine (CIBM) program (funded by the National Library of Medicine (NLM)) and for giving me the opportunity to collaborate on his research related to Schwann cells. Professor Svaren's friendliness, enthusiasm, and readiness to provide feedback are deeply appreciated. His biological insights were invaluable to my work, and I cherish the opportunity I had to present our collaborative research at the American Society for Human Genetics (ASHG) conference in Los Angeles, California. I am also deeply thankful to Professor Mark Craven for his pivotal role in my acceptance into the U.W.-Madison graduate school and the CIBM program. I want to thank him for serving as the chair of my Ph.D. committee. I want to thank Professor Qiongshi Lu for his insightful instruction in Statistical Genetics that was useful for my research. My conversations with Professor Craven and Professor Lu during the Open House for accepted students in the Biomedical Data Science Ph.D. program were key factors in my decision to pursue my Ph.D. at U.W.-Madison.

To the Daifeng Wang lab at the Waisman Center and the CIBM program, I am deeply thankful for the camaraderie, team spirit, and unwavering support throughout the years. The lab has been more than just a workplace; it has been a family where we celebrated each other's achievements, engaged in thought-provoking discussions, and supported one another through

challenges. We have formed good friendship bonds, and I will cherish the warmth and unity that made even the cold winters in Wisconsin bearable. For the SNPheno paper, I appreciate my Undergraduate Research Scholar student and mentee, Jonathan Edward Bryan, for his assistance and valuable feedback. In the development of the NetREm methodology, I am thankful to my co-authors Dr. Xiang Huang, Dr. Raghu Ramesh, Professor John Svaren, and Professor Daifeng Wang for their direct contributions. I also acknowledge the insightful feedback and guidance from Dr. Kalpana Hanthanan Arachchilage, Noah Cohen Kalafut, Sean Chang, and Harrison Pantera. My appreciation also extends to Sayali Anil Alatkar, Dr. Ting Jin, Dr. Chirag Gupta, Dr. Pramod Bharadwaj Chandrashekar, Jie Sheng, and Dr. Shuang Liu., for their friendship, kindness, and support.

I am profoundly grateful to my family for their unwavering love, support, and encouragement throughout this journey. Each of them, in their own unique way, has stood by me through every trial and triumph, providing strength when I needed it most. Words cannot fully capture my gratitude to my parents: my mother, Renu (Poochie) Khullar, and my father, Sanjeev Khullar (a proud alumnus of U.W.-Madison's Master's in Computer Science program). Their constant encouragement and belief in me have been my foundation. Their visit to U.W.-Madison during the accepted students' event was a pivotal moment in my decision to pursue this program, and their blessings have been my guiding force. I am especially fortunate to have Poochie's unwavering support as both my best friend and mentor during the crucial years of my Ph.D., shaping this journey in ways I will forever cherish. I hope to live up to Poochie's faith in me to scale my computational tools (SNPheno and NetREm) to newer heights post-PhD, and I have made a commitment to do so. I am also incredibly fortunate to have the support of my older sister, Avantika Khullar Argolo (Didi), and younger brother, Armaan (Baba) Khullar. My family continued to grow during my Ph.D. journey, and I am blessed to be cheered on by my brother-in-law, Felipe Argolo; my niece and nephew, Sofía Leila Argolo and Lucas Henrique Argolo; and their quintessentially British cat, Bandit. Becoming a cat mom during this time has been another joy. Tubby, my affectionate and very vocal orange tabby, who I adopted from the Dane County Humane Society in Wisconsin, and Chia, my rescue from Utah with her loyal nature and beautiful blue eyes, have brought warmth, love, and companionship to my life. I also wish to honor the memory of my late maternal grandmother, Dr. Raj Kumari Bhandari (internal medicine physician), who instilled in me a love of learning and science. My last conversation with her was a promise that her little Chanu Manu (me) would become a doctor and make a difference in public health like she did. This dissertation is a step toward fulfilling that promise. I am also thankful to my paternal grandmother: Santosh Khullar (Dadima) and maternal aunt: Nina Thakur (Maasi) for their love and prayers from across the miles.

My academic journey has been profoundly shaped by my alma mater, Georgetown University. The Jesuit values of *Cura Personalis* (care for the whole person) instilled in me a deep desire to help my community and make a difference. I owe much to the professors who encouraged me to use my quantitative skills for social good, starting with Professor Oded Meyer, who sparked my interest in biostatistics and encouraged me to attend the Summer Institute in Biostatistics at the University of Iowa College of Public Health. This experience was pivotal in steering me toward further research in this field. Professors David Caraballo, Hans Engler, Canan Ulu, Keith Ord, and Alisa Carse are among many educators at the Hilltop who shaped my path. *Hoya Saxa!* (Latin: What rocks!). I am grateful to healthcare startup DecisionQ, for giving me an opportunity to collaborate on a project leading to my first publication. I am also thankful to the individuals I encountered while working as a nursing assistant and as a volunteer in hospitals, hospices, nursing homes, schools, and villages in developing countries like Guatemala and Ecuador. Their stories inspire me to persevere and use my skills to make a difference in healthcare.

During my Ph.D., I received invaluable support from Dr. Sara Knaack, Dr. Sean McIlwain, Sonja Oetzel, Professor Irene Ong, Professor Mouna Ayari Ben Hadj Kacem, Professor Tyler R. Caraza-Harter, Professor Sushmita Roy, Professor Christina Kendziorski, Professor Michael Newton. My heartfelt thanks also go to Shelley Maxted and Beth Bierman from the Biomedical Data Science program and Louise Pape from the CIBM program for their vital administrative support over the years. I am thankful to the Waisman Center and U.W.-Madison for their support over the years. Through U.W.-Madison, I expanded my knowledge in Computer Science, Genetics, Neurogenetics, Immunology, Molecular Biology, and Biostatistics, obtaining a Master's in Computer Science along the way. And, at U.W.-Madison, I had the joy of attending my 1st college football game. Go *Badgers*! My Madison community is one of the best gifts. I am deeply thankful for all the love, kindness, and support from my family across the United States of America, United Kingdom, India, and beyond. Their unwavering belief in me has been my strength throughout this journey. I am also profoundly indebted to my dear friends who come from all walks of life and represent the different parts of my past. They gave me support and I am grateful to have their faith in me during this journey.

Table of Contents

§ Chapter 1: Introduction

§ 1.1 Background and Motivation

The human brain, a complex central hub, orchestrates multitudes of functions – from governing organs and tissues to shaping behaviors, emotions, intelligence, thoughts, memories. These aspects collectively define our humanity and personality (NINDS NIH et. al 2013). Recent statistics highlight the gravity of neurological disorders, the leading cause of death in 2016 (GBD 2016 Neurology Collaborators et. al 2019). Moreover, nearly 20% of U.S. adults were affected by mental illnesses in 2019 (NIMH NIH et. al), and these statistics are worsening and becoming bleaker. For instance, over 33% of Covid-19 survivors have developed psychiatric or neurological illnesses (e.g. Alzheimer's disease (AD)) within 6 months post-infection (Nania et. al 2021). Tackling fundamental questions in this domain is crucial for advancing precision medicine and helping pave a way for developing targeted therapies and interventions globally. During my Ph.D. program, I have focused on researching and helping advance our understanding of challenges related to *4 key research gaps* in this area. In this section, I provide background and motivation for these 4 problems and highlight existing approaches and limitations in addressing them.

Patient population data on cohorts with various brain diseases measure several phenotypes like disease progression. Many complex diseases exist on a clinical continuum, with some patients having milder and others having debilitating intensity levels (Ivleva et al. 2010). Identifying neurocognitive phenotypes may help explain the underlying biology of complex brain disorders (Congdon et al. 2010), as the extent of impairment varies greatly. "There is no one type of autism, but many" (AutismSpeaks.org), as Autism Spectrum Disorder (ASD) exists on a spectrum of degrees of difficulty with social skills, speech, repetitive behaviors, non-verbal communication, educational attainment (ASD severity endophenotype (Wong et. al 2021)). With neuronal connections and neurons dying, AD patients gradually forget aspects about themselves, losing memory, cognitive abilities, executive function. Many underlying molecular changes happen (e.g. amyloid-β plaques, chronic neuroinflammation, neurofibrillary tangles (NFTs)). AD phenotypes include stages, MMSE scores (measure cognitive function/awareness in elderly), NFTs, Braak

progression (neuropathology), CERAD scores (Blalock et al. 2011), diagnosis age, cognitive diagnosis, age at death (Bennett et al. 2018), cognitive resilience (ability of individuals with neurodegeneration to ward off AD (Aiello Bowles et al. 2019)). A study on Schizophrenia (SCZ) populations analyzed stages (clinical high risk for psychosis, first episode of psychosis, chronic SCZ (del Re et al. 2015)) while another focused on 2 molecular SCZ subtypes (Bowen et al. 2019). A Multiple Sclerosis (MS) study looked at 3 categories based on frequency of MS-related attacks, recovery states, neurological deficits, impairment levels (Tajouri et al. 2007). Another study (Cappelletti et al. 2023) measured Parkinson's disease (PD) phenotypes related to the degree of Lewy bodies (LBs, i.e. Alpha-synuclein protein aggregates in different brain regions) in the frontal cortex of 84 postmortem donors (23 healthy, 61 with varying degrees of LB pathology). Phenotypes (e.g. activity level count, time in an elevated zero-maze, percent of time spent in open quadrants) helped study molecular interplays among stress, alcohol, and anxiety in mice Hippocampal tissues (Luo et al. 2018). Electronic Health Records (EHRs) are also used to obtain disease phenotypes (Strauss et al. 2021). Properly defining Bipolar Disorder (BD) phenotypes, using diagnosis and clinical features, has enabled genetic investigations to uncover BD susceptibility genes (MacQueen et. al 2005). These diseases present a wide clinical spectrum, with diverse phenotypes related to their onset, progression, severity.

Adult humans have over 37 trillion cells, categorized into ~200 distinct types (Bianconi et al. 2013). Each cell type, despite having identical DNA, expresses a unique set of genes relevant to its specific role and context in the body. Abnormal gene expression is often linked to diseases. Investigating these gene expression patterns can improve our understanding of cell-type-specific disease mechanisms and may help identify key genes involved in diseases and their associated phenotypes (Wong et. al 2021).

Emerging single-cell data analyses have shown many target genes (TGs) have cell-type-specific gene expression patterns, highlighting gene regulation is both cell-type and context-dependent. The composition of brain regions, which varies in diseases such as AD (e.g. potential increase in immune cells and decrease in neurons), can be effectively studied using single-cell data like scRNA-seq (i.e. single-cell gene expression), scATAC-seq, DNase-seq, scChIP-seq (Jiang et. al 2020). This data can help reveal

intricate TG dynamics in complex biological processes across various cell types (He et al. 2023), helping identify genuine biological variants (e.g. single nucleotide polymorphisms (SNPs)), disease-associated TGs, specific cell types pathologically targeted by diseases (Sealfon et. al 2021). SNPs are the most common type of genetic variation, where a single base change in the DNA sequence can influence gene function, contribute to individual differences, and impact disease susceptibility. Techniques on single-cell data allow for exploration of functional genomics, epigenetic signatures, regulatory mechanisms, and gene expression at a cell-type level, capturing the complete transcriptome of individual cells, shedding light on intricate characteristics of various tissues. Through scRNA-seq, cells from different individuals can be analyzed simultaneously, measuring thousands of genes per cell (Wang et al. 2021b), thereby mitigating technical batch effects, enabling detection of expressed cell-type-specific genes. Recent advancements in single-cell multi-omics can revolutionize our understanding of cell-type-specific TG regulation in brain diseases.

Nonetheless, understanding gene regulation, especially in the context of cell-type-specific gene expression and deeper disease phenotypes in brain disorders, is still challenging (**Gap #1**). The regulation of transcription of a given TG involves various regulatory factors, including non-coding SNPs, proteins like Transcription Factors (TFs), regulatory elements (e.g., enhancers, promoters). These factors collectively form gene regulatory networks (GRNs), comprising directed regulatory relationships from TFs to TGs, which respond to extracellular signals to control gene expression and functions, determining cell types, identities, disease states (Sinha et. al 2020). GRNs are pivotal in transcriptional regulation, helping explain how TFs bind to specific TF Binding Sites (TFBSs) on regulatory elements to regulate transcription and subsequent expression of their TGs. State-of-the-art tools (e.g. BEELINE (Pratapa et. al 2020), SCENIC (Aibar et. al 2017), GRNBoost (Moerman et al. 2019), Signac (Stuart et. al 2021), scGRNom (Jin et al. 2021), PoLoBag (Roy et. al 2020), TIGRESS (Haury et al. 2012), SCODE (Matsumoto et al. 2017), Inferelator 3.0 (Skok Gibbs et al. 2022), CellOracle (Kamimoto et al. 2023)) build cell-type GRNs from single-cell gene expression data, using co-expression, correlation, differential equation, Bayesian-network, information-theoretic, machine learning, and/or multi-omics integration-based techniques to

infer potential relations among individual candidate TFs and their TGs that they help regulate (e.g. TF-TG regulatory links). Nonetheless, gene regulation in eukaryotic organisms, like humans, is inherently intricate; these tools may discard TF-TG pairs with weak or de-coupled relations (e.g. uncorrelated expression), when such TFs could instead regulate the TG through more complicated avenues (e.g. joint coordination with other TFs)(Zaborowski and Walther 2020). These tools may also struggle to account for the noisiness of gene expression data, which may also be attributed to coordination among TFs (Parab et al. 2022). Further, there is an issue of high correlation among TF predictors, and these tools may drop some of these correlated predictors (that may include true, causal TFs for the TG regulation) and/or opt for more independent predictors (that may not necessarily be causal). Thus, a core drawback of these tools is that they mainly consider TFs in isolation, overlooking the interdependent protein-protein interactions (PPIs) among coordinating TFs that are critical for cell-type TG regulation.

Complicating gene regulatory mechanisms is this fact that, in humans, combinations of TFs often interact directly or indirectly (typically cooperate not compete) along with the transcription initiation complex to coordinate regulation of their TGs. The extent and implications of this TF coordination, driving TG regulation genome-wide, is a widely observed yet poorly understood phenomena (Ibarra et. al 2020), especially across cell types. Uncovering the interactome (set of molecular physical interactions among biological entities may help explain how gene functions and regulation work together; a significant hurdle in network biology is the quality and coverage of interactome data of PPIs (Sevimoglu and Arga 2014). These protein interactions are involved in orchestrating various biological processes among organelles and structures in the cell (Liu et al. 2018). While SCINET (Li and Li 2008) reconstructs cell-type interactomes, it falls short in illustrating TG regulation. TF-Cluster (Nie et al. 2011) identifies functionally coordinated TFs involved in biological processes but does not focus on TG regulation and is based on coexpression analysis (and does not use any existing prior knowledge). That is, there is this knowledge gap in understanding the TF-TF coordination networks of indirect and direct interactions among TFs at the protein-level that are involved in the regulation of TGs (**Gap #2**). Current human PPI networks of direct and/or indirect functional associations among TFs are often at a global level (i.e. are

not cell-type-specific (Yeger-Lotem and Sharan 2015)) and may not encompass all interactions among TFs that are important for cell-type gene regulation. These PPI networks are largely incomplete, hindering the study of network properties of disease genes. Estimates suggest that there may be ≈130-650k human PPIs, but only a fraction of them are identified through experiments (Sevimoglu and Arga 2014; Venkatesan et al. 2009) and may include incomplete and false positive (FP; false PPI) links (Yu et al. 2020).

This lack of cell-type-specific PPI networks presents another challenge to be tackled. On-going research efforts urge developments in bioinformatics and computational biology tools to cost-effectively help annotate existing PPIs and discover and nominate novel PPIs (to be examined by follow-ups in wet-lab studies) at the cell-type-level (Yu et al. 2020). There is a growing need for condition-specific (e.g. cell-type) interactomes that represent protein interactions in specific tissues or under conditions, which can enhance the significance of network analysis, especially when studying disease-associated alterations in PPIs. Brain disorders like AD demonstrate tissue-specific gene expression changes attributed to transcriptional regulation by TFs (Pearl et al. 2019), which likely coordinate with each other in TF-TF coordination networks. Risk variants (e.g. SNPs) for diseases (e.g. developmental disorders) are known to alter PPIs. Nonetheless, the specific impact of these mutations on not only diseases but also on disease-related phenotypes is largely unexplored (Cheng et al. 2021), and efforts to annotate PPIs at a cell-type level are in their infancy.

In light of these varying phenotypes in brain diseases, researchers are exploring how genetic variants (e.g. SNPs) influence disease genes and phenotypes, as that is still unclear. Genome-Wide Association Studies (GWAS) have identified numerous SNPs associated with AD and other brain disorders, establishing Disease-SNP links. However, these SNPs typically exhibit weak effect sizes and contribute cumulatively to the complexity of these diseases. Remarkably, ~90% of these SNPs are on non-coding DNA (Maurano et. al 2012; Li and Ritchie 2021), often within regulatory elements like TFBSs. These locations suggest they might affect TF binding, consequently disrupting TG regulation and expression, either locally or distally (Farrow et al. 2022). Such disruptions can influence core disease genes and phenotypes within complex networks, underscoring the need to explore their dysregulation effects on

disease outcomes. Nonetheless, deciphering the molecular and cellular impact of SNPs on TG expression in brain diseases represents a formidable challenge. While computational tools for analyzing the impact of coding SNPs on disease outcomes are relatively advanced, frameworks for studying non-coding SNPs are still budding (Wong et. al 2021; Novikova et al. 2021a). Disease GWAS, when combined with expression quantitative trait loci (eQTL: SNP-TG links), can identify disease risk SNPs associated with TG expression; phenome-wide association studies (PheWAS) test for meaningful associations between GWAS SNPs and disease phenotypes (Diogo et al. 2018). Yet, these approaches often fall short in elucidating the detailed gene regulatory mechanisms by which these SNPs influence disease-related TGs and their respective role in deeper disease phenotypes, especially as these diseases exist on a spectrum (**Gap #3**). This gap highlights a broader challenge in the field (Sevimoglu and Arga 2014): effectively integrating diverse data sources to understand how genetics (including non-coding mutations) and environmental factors impact disease phenotypes functionally. Thus, developing methods to understand the impact of non-coding SNPs on disease phenotypes, particularly at the cell-type and tissue levels, is vital for uncovering novel genetic mechanisms associated with disease phenotypes. This is crucial, given the significant role of regulatory variation in contributing to phenotypic diversity observed in human populations (Thompson et. al 2015).

Researchers utilize the deeper population phenotypes for complex diseases along with gene expression data to perform various analyses. For instance, they employ Differential Expression (DE) analysis to identify significant DE genes (DEGs) (Love et al. 2023) associated with phenotypes and co-expression networks (e.g. Weighted Gene Co-Expression Network Analysis (WGCNA) (Langfelder and Horvath 2008) on bulk data, scWGCNA on single-cell data (Morabito et al. 2021)) to correlate genes with gene modules, disease phenotypes, biological pathways, enrichments. Co-expressed genes typically share similar expression dynamic patterns, implying that they may share a common set of TFs binding to their regulatory elements (i.e. are likely to be co-regulated (Allocco et al. 2004)), ensuring their coordinated transcription (van Duin et al. 2023). Still, co-expression network analysis is unsupervised learning and learns several False Positive associations (Badia-i-Mompel et al. 2023). Further, these computational

methods do not adequately explain patterns of TG regulation and direct regulatory interactions. It is vital to also uncover underlying gene regulatory mechanisms controlling DEGs, co-expression networks, gene expression dynamics (Pearl et al. 2019). There is a need for integrative methods to build comprehensive GRNs (e.g. combining GRNs and gene co-expression networks) to help uncover how perturbations in GRNs may be associated with changes in disease phenotypes and other context-specific biological pathways. GRNs are entangled in cellular machinery and it is vital to link gene regulation to these cellular processes (Badia-i-Mompel et al. 2023).

Machine Learning (ML) approaches have significantly enhanced the integration of multi-omics data, improving predictions about the impact of SNPs on phenotypes, the effects of dysregulated TGs on biological pathways, and future disease risks at both tissue and cell-type levels (Sealfon et. al 2021). There is increasing focus on network-based ML methods, which connect non-coding SNPs with potential TGs and use network biology to identify novel disease-related TGs (Wong et. al 2021). An example is NetWAS (Network-Wide Association Study), which uses tissue- and cell-type-specific networks along with gene-level P-values in an ML classifier to re-prioritize GWAS associations, thereby identifying TGs associated with diseases (Wong et al. 2018). Such methods incorporate the latest multi-omics data across various tissues and diseases, revealing insights into the associations between SNPs, and control versus disease states. Other studies have tried to integrate co-expression networks with edge-weighted PPI networks to construct disease-specific networks to uncover biomarkers for diseases (Su et al. 2022). However, these current ML approaches (e.g. NetWAS) are limited in their ability to fully explore prioritized GRNs for disease and phenotype prediction, especially at the cell-type level. Multi-omics data, combined with networks on gene regulatory mechanisms and pathway ML models, are crucial for deepening our understanding of human diseases (Wong et. al 2021).

Recent literature underscores the need for predictive models that include information on genes, modules, and pathways involved in diseases, focusing on how network perturbations affect disease risk (Wong et. al 2021; Chandrashekar et. al. 2023). Network biology advocates the construction of ‘triggers networks’, integrative biological networks (weaving networks related to: regulatory interactions, PPIs,

metabolism, signaling) that relate potential biomarkers to disease mechanisms (Sevimoglu and Arga 2014). While many computational methods integrating multi-omics data have been developed for predicting brain disease biomarkers, they often fall short in fully elucidating the link between genotype and phenotype through gene regulatory mechanisms at both tissue and cell-type levels (Chandrashekar et. al. 2023; Gupta et. al 2022). Designing network-based predictive ML models that incorporate biological ground truth data, including regulatory, interaction, and functional networks associated with diseases, is key to predicting disease phenotypes and identifying candidate biomarkers (**Gap #4**).

§ 1.2 Research Contributions

§ 1.2.1 Overview

In response to the 4 broad challenges highlighted in **§1.1**, I have developed 2 innovative computational tools, **SNPheno (§ Chapter 2, Figure 1.1)** and **NetREm (§ Chapter 3, Figure 1.2)**.

- SNPheno: Predict the role of non-coding SNPs on gene regulatory mechanisms and phenotypes
- NetREm: Network Regression Embeddings reveal cell-type TF coordination for gene regulation

My research uses multi-omics data to investigate gene regulatory mechanisms (e.g. GRNs, gene co-expression networks, TF-TF coordination networks) and their links to brain disease phenotypes, aiming to create a comprehensive narrative from genotype to phenotype. I built SNP-effected-GRNs to elucidate how non-coding disease SNPs may potentially affect complex biological processes and contribute to disease severity. Further, my work examines the evolution and variability of regulatory and TF coordination mechanisms across brain regions and cell types during disease progression. This includes identifying cell-type-specific PPI subnetworks that play roles in TG regulation.

Many of the networks that I construct (e.g. TF-TF coordination networks, SNP-effected-GRNs, comprehensive GRNs) are unprecedented. Hence, creative methods are needed to utilize their knowledge and data structures to make inferences on potential biomarkers. Thus, I apply ML models to these networks to identify potential biomarker SNPs and TGs for brain diseases. In summary, **SNPheno** and

NetREm are poised to help deepen our understanding of brain-related diseases and aid in identifying network-based biomarkers for disease phenotypes.

In my personal capacity, I regard the development of the SNPheno and NetREm computational tools as my most significant contributions to the field thus far. These tools have been pivotal in advancing our understanding of genetic regulation and disease mechanisms, and I see them as foundational building blocks for further research, laying the groundwork for future innovations in precision healthcare and personalized medicine. As I write about these tools, I do so with both humility and pride.

§ 1.2.2 Computational tools

SNPheno (Khullar and Wang 2023, 2021): addresses Challenges 1, 3, and 4. SNPheno is a novel computational method to integrate multi-modal features at the cell-type level to link non-coding SNPs to disease genes and phenotypes for various brain diseases. It is challenging to identify genetic variants (e.g. SNPs) that may alter GRNs in brain disorders. An integrated annotation at a base-pair resolution, across cell and tissue types, is required for mechanistic and diagnostic predictions and determining shared/different mechanisms among these diseases. Through SNPheno we can better understand the molecular architecture of various brain diseases, namely both the: shared regulatory mechanisms and differential disease, epigenetic, tissue, and cell-type/tissue-type specific impacts of these SNPs across complex diseases. SNPheno networks help better explain the role of non-coding SNPs on disease phenotypes at a cell and tissue-type level to provide novel insights on genetic mechanisms associated with brain-related disease phenotypes.

NetREm (Khullar et al. 2023): addresses Challenges 1, 2, 3 (partially), and 4. NetREm is an innovative computational method that integrates multimodal data to output cell-type-specific regulatory and coordination networks that predict how TFs coordinate (cooperatively or antagonistically) in groups to co-regulate TGs. TFs act in concert with each other in biologically-meaningful PPIs to regulate target gene expression. Direct as well as indirect TF-TF interactions may be captured by PPI networks (PPINs). Traditional, state-of-the-art GRN-inference tools do not consider PPIs when predicting TF-TG relationships. NetREm is designed to improve upon weaknesses in these existing tools that do not

consider the interdependent role of TF coordination with other TFs for regulating TGs; it leverages network-regularized regression to predict expression of target genes subject to a PPIN constraint among the predictors. Overall, NetREm inputs gene expression, a TF-TF PPIN, and an optional prior state-of-the-art cell-type GRN that links TFs to target genes TGs (initial feature selection of TFs).

§ 1.3 Research Outline

In **Chapter 2**, I introduce SNPheno, a computational pipeline to connect disease-associated SNPs, GRNs, and TGs implicated in complex phenotypes. This tool aims to shed light on the influence of SNPs on transcriptional dysregulation by examining the effect of alterations in TFBSs on TF regulation of TGs, disease phenotypes, abnormalities in biological pathways. SNPheno creates a comprehensive SNP-effected-GRN linking disease SNPs to their respective TGs, phenotypes, pathways. In this chapter, we apply SNPheno to AD and Covid-19, focusing on 3 brain regions dysregulated in AD. This analysis uses bulk-level data comprised of diverse cell types. Initial steps include gene co-expression network analysis in these regions, revealing that certain genes exhibit similar expression dynamics in relation to AD progression and associate with rogue immunity. This underscores interactions among TG expression, neuroimmunology, Covid, AD. Subsequent GRN predictions identify TFs governing co-expressed genes. By examining GRNs and AD-phenotype-related co-expression modules and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways for AD and Covid, we identify AD-Covid genes in each brain region as primary features in our ML analysis for predicting Covid severity. We identify 36 predictive AD-Covid genes across our models and use them to differentiate AD from Controls in another brain region (Zhang et al. 2024). Notably, our ML performance in predicting Covid severity and AD surpasses existing methods using established marker genes. Finally, we map AD and severe Covid GWAS SNPs onto our GRNs to explore functional mechanisms and AD phenotypes related to these SNPs, particularly focusing on the 36 predictive AD-Covid genes.

In **Chapter 3**, I introduce NetREm (Khullar et al. 2023) a novel method using network regression embeddings to reveal complex networks of TF coordination that regulate TG expression. To provide a more detailed representation of cell-type-specific TF-TG regulatory networks and TF-TF coordination

networks, NetREm incorporates additional data, such as PPIs. NetREm uncovers cell-type-specific TF-TF PPI subnetworks and novel cell-type TF-TF links for follow-ups to examine (i.e. if both TFs in that TF-TF link interact directly or indirectly). That is, NetREm identifies functionally-coordinating groups of TFs, likely involved in biologically meaningful PPIs, which may co-regulate the TGs. I present simulation studies on toy data to showcase NetREm's methodology and its relative advantages. Then, I benchmark the performance of NetREm's TF-TG regulatory networks and TF-TF coordination networks in various cell-types spanning humans and mice. Further, I utilize multiomics data to derive initial GRNs of candidate TF-regulatory element-TG links, which I input to NetREm for 2 real-world applications in humans: 1. myelinating (mSCs) versus non-myelinating (nmSCs) Schwann cells (SCs); 2. Eight glial and/or neuronal cell-types in AD versus Controls. In both applications, we apply NetREm to construct both TF-TG regulatory networks (complementary to a GRN) and unprecedented TF-TF coordination networks. These initial GRNs are constructed using various multi-omics data (e.g. scATAC-seq data on accessible DNA regions for the cell-type, cell-type cis-candidate regulatory elements (cCREs) of chromatin interactions, eQTL data, gene expression data, PPIs, TF binding profiles). We validate top findings using available functional genomic data in humans, rats, or mice (e.g. eQTLs, GWAS, CUT&RUN sequencing, ChIP-seq data, knockout studies). We obtain TF-TF links annotated with neurodegenerative diseases. Then, we compute changes in cell-type TF-TF coordination network embedding values between Control and AD across TGs as input to our ML models; these models identify candidate cell-type biomarker TGs for neurodegeneration. Further, we present examples of GWAS-eQTL colocalized disease risk SNPs altering Transcriptional Regulatory Modules (TRMs, subnetworks of TF-TF coordination networks) and TF-TG regulatory networks associated with disease-associated TGs.

In **Chapter 4**, I summarize and critically assess the limitations of SNPheno and NetREm, discussing the implications of both methods and outlining potential avenues for future research. Please note that supplementary Chapter A (SNPheno) and Chapter B (NetREm) are publicly-available on GitHub (SaniyaKhullar 2024) and contain supplementary methods and materials, figures, tables, data files, and hyperlinks for Chapters 2 and 3 of this main dissertation, respectively. Proteins (e.g. TFs, and TGs

(*italics*) are represented by HGNC (Human Genome Organization (HUGO) Gene Nomenclature Committee) symbols. It is my hope that the insights and discoveries from this dissertation will pave the way for future innovations and interventions that can transform patient care and public health outcomes.

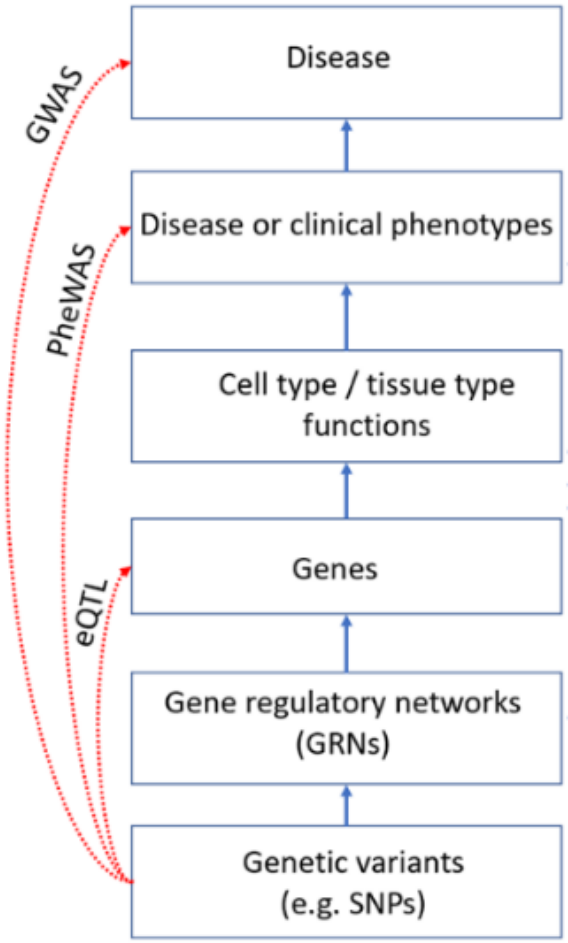


Figure 1.1 – Challenges addressed by SNPheno

Current approaches do not explain the entire gene regulatory mechanisms that link SNP to disease outcomes or phenotypes. It is important to understand the full regulatory mechanisms. For instance, GWAS associate SNPs with diseases. PheWAS associate SNPs with disease phenotypes, and eQTL data tries to link expression quantitative trait SNPs (i.e. eSNPs) with changes in target gene expression (i.e. eTGs) in eSNP-eTG pairs. SNPheno helps integrate this in a holistic method.

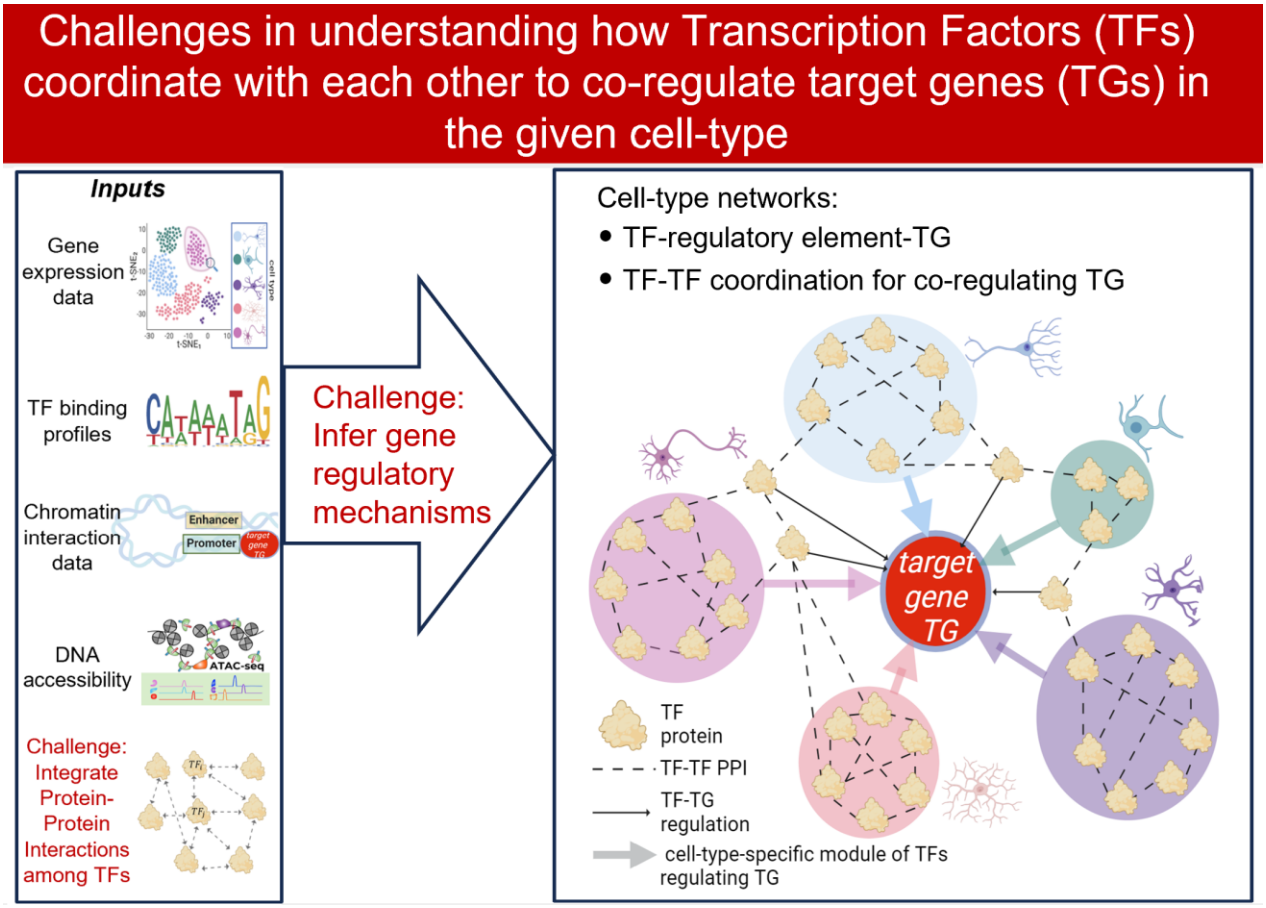


Figure 1.2 – Challenges addressed by NetREm

It is challenging to integrate PPIs among TFs along with other multi-omics data to infer gene regulatory mechanisms at the cell-type and tissue-type levels.

§ Chapter 2: SNPheno: Predicting brain-regional gene regulatory networks from multi-omics for Alzheimer's disease phenotypes and Covid-19 severity

§ 2.0 Abstract

Neuroinflammation and immune dysregulation play a key role in Alzheimer's disease (AD) and are also associated with severe Covid-19 and neurological symptoms. Also, genome-wide association studies (GWAS) found many risk single nucleotide polymorphisms (SNPs) for AD and Covid-19. However, our understanding of underlying gene regulatory mechanisms from risk SNPs to AD, Covid-19 and phenotypes is still limited. To this end, we performed an integrative multi-omics analysis to predict gene regulatory networks for major brain regions from population data in AD. Our networks linked transcription factors (TFs) to TF binding sites (TFBSs) on regulatory elements to target genes (TGs). Comparative network analyses revealed cross-region-conserved and region-specific regulatory networks, in which many immunological genes are present. Furthermore, we identified a list of AD–Covid genes using our networks involving known AD and Covid-19 genes. Our machine learning analysis prioritized 36 AD–Covid candidate genes for predicting Covid severity. Our independent validation analyses found that these genes outperform known genes for classifying Covid-19 severity and AD. Finally, we mapped genome-wide association study SNPs of AD and severe Covid that interrupt TFBSs on our regulatory networks, revealing potential mechanistic insights of those disease risk variants. Our analyses and results are open-source available, providing an AD–Covid functional genomic resource at the brain region level.

§ 2.1 Introduction

AD, a neurodegenerative disease, affects over 50 million elders worldwide (Alzheimers.net). Late-onset AD (LOAD) comprises > 97% of all AD cases, usually occurring after age 65 (Rabinovici 2019). AD patients experience phenotypic changes such as memory loss, cognitive decline, weak executive function (Alzheimers.net) (e.g. poor Mini-Mental State Exam (MMSE) scores). Many underlying molecular changes happen like an accumulation of amyloid-beta (Aβ) plaques and neurofibrillary tangles (NFTs), chronic

neuroinflammation (may begin decades before clinical onset). Nonetheless, molecular mechanisms behind AD progression and phenotypes remain elusive. Misguided innate immunity may be a major culprit driving AD based on the neuroimmunomodulation theory of AD (Maccioni et. al 2018).

Molecular interconnections that exist between the central nervous system (CNS) and immune system (Zass et al. 2017) are also seen via the strong correlations between AD and the severity of Covid-19 infection (Thompson et. al 2022). Covid-19, a robust marker for an overreactive immune system, can also mediate neuroinflammation (Amruta et. al 2021). β-coronaviruses (like Covid-19) may attack the CNS, elevating AD dementia processes (Erausquin et. al 2021). Covid survivors have greater risk of neurological/psychiatric problems and brain fog (neuro-Covid (Heming et al. 2021) or long-Covid); patients may have visible neuropathological abnormalities in brain structure (Thompson et. al 2022) (e.g. Hippocampus atrophy (Gordon et. al 2021)) similar to those found in AD patients (Reiken et. al 2022). AD brains have high levels of circulating pro-inflammatory cytokines associated with activation of microglia [macrophage resident immune cells typically downregulated in healthy brains (Zass et al. 2017); these cytokines also contribute to the cytokine storm causing exaggerated inflammation characteristic of severe Covid (Su et. al 2021). In fact, Covid patients experiencing delirium (symptom linked with high risk of AD) are at grave risk of death and typically sent to an Intensive Care Unit (ICU) (Gordon et. al 2021). Elder patients (age > 65 years) are 70% likelier to be diagnosed with AD within a year of Covid infection (Lindsey et. al 2022). There is a 2-fold increased risk of Covid death in AD patients (Anderson et. al 2021) and of higher severity of Covid for patients with APOE4 (E4 alleles for the key LOAD risk gene APOE) (Inal 2020). Links among Covid, cognitive decline and neurodegenerative diseases like AD are puzzling and poorly understood (Gordon et. al 2021). AD itself has over 34 canonical and intricately interconnected pathways, making that process daunting (Mizuno et. al 2012). Focusing on AD–Covid pathways may be a useful starting point of departure given their strong links. Thus, understanding genetic effects and underlying molecular mechanisms for shared AD–Covid paths may shed more insights on rogue immune responses not only in AD but also in severe and/or neuro Covid-19.

Several neuroimmunology pathways are shared by AD and Covid. One of them is the NF-κB (Nuclear Factor Kappa-light-chain-enhancer of activated B cells) pathway that is found in almost all cell-types and that regulates, *inter alia,* brain homeostasis (maintains synapse plasticity, learning, memory; moderates neuron survival/apoptosis)(Jha et. al 2019), innate immunity, inflammation (Lawrence 2009). A prominent hypothesis believes AD may be caused by an impaired NF-κB pathway (Jha et. al 2019) with overactivated NF-κB transcription factors (TFs) like RELA and NFKB1. This may lead to more cytokines, neuroinflammation, oxidative stress complications, activated microglia, neuron death (Jha et. al 2019). NF-κB TFs are also involved in a positive feedback loop, activating pro-inflammatory cytokines in severe Covid (Su et. al 2021). RELA, one of the most important TFs regulating Covid response (Fagone et. al 2020), is associated with APOE4 (Xiong et al. 2021). Gene regulatory networks (GRNs) can capture how these TFs regulate several genes of pro-inflammatory cytokines. Thus, to understand neuroimmunology in Covid-19 and AD better, it is important to analyze these underlying gene regulatory mechanisms.

Gene expression and regulation are key mechanisms leading to human diseases. Studies found differentially expressed genes (DEGs) in AD in various brain regions like the Hippocampus Cornu Ammonis 1 (CA1), Lateral Temporal Lobe (LTL), Dorsolateral Prefrontal Cortex (DLPFC). In particular, the CA1 region—which is crucial for autobiographical memory, mental time travel, and self-awareness—usually has the biggest loss in memory ability, neurogenesis, volume and neuronal density in the AD Hippocampus (Bartsch et. al, 2011). The LTL contains the cerebral cortex (responsible for hearing, understanding language, visual processing, and facial recognition) (Goldstein et al. 2017) and is impacted early in AD (Nativio et. al 2020). The DLPFC is involved in executive functioning (working memory and selective attention), supports cognitive responses to sensory information (Sturm et al. 2016), works with the Hippocampus to help mediate complex cognitive functions (Brinton et al. 2009), and has plasticity deficits in AD patients (Kumar et. al, 2020). It is still challenging to understand the molecular and cellular mechanisms that fundamentally drive the early progression of AD, especially in these three brain regions.

Gene co-expression networks are widely used to identify co-expressed gene modules and link expression patterns to AD phenotypes (Morabito et. al 2020). Genes in a module show similar expression

dynamics across phenotypes, denoting that they share certain molecular mechanisms that are dysregulated in AD (Wan et al. 2018). Nevertheless, understanding gene regulatory mechanisms controlling DEGs, co-expressed genes and modules for various AD phenotypes as they relate to the immune system is unclear. Gene expression and function are controlled by various regulatory factors working together in a GRN, like: TFs binding to TF binding sites (TFBSs) on regulatory elements (e.g. enhancers, promoters). However, our understanding of gene regulation in AD and in AD–Covid is still limited.

Researchers have identified genetic risk variants associated with various brain-related diseases like AD and found that most risk SNPs (occur at a single DNA position among individuals) may be involved in changes in gene regulation. Over 90% of disease risk SNPs from GWAS are in non-coding regions (Kumar et al. 2017), as 98.5% of human DNA is non-coding. GWAS SNPs shed more light on specific biological effects of certain variants and mechanisms associated with complex disease phenotypes. This genetic variation can lead to differential disease risk in people; harmful protein-coding SNPs typically lead to downstream effects (e.g. truncated, loss-of-function, or harmful proteins; altered protein properties or structures) that can be properly examined (Wong et. al 2021). GWAS linked LOAD SNPs to genes by proximity to coding DNA(Novikova et al. 2021b), uncovering risk genes associated with microglia, increased cytokines, activation, neuroinflammation, worse AD (Kinney et al. 2018a); besides APOE4, no causal AD determinants are known. This is the "missing heritability problem": GWAS SNPs explain just a small fraction of total heritability of complex diseases (Wong et. al 2021).

Most brain-related diseases like AD are polygenic, shaped by many SNPs, genes, environmental factors (e.g. epigenetics) that are still poorly understood (van der Wijst et al. 2020). Recent studies support an omni-genic model, where complex diseases can be attributable to thousands of harmful SNPs battling other protective SNPs (Wong et. al 2021), affecting genes that can further impact core genes (via interactions, shared functions, pathways, networks). Thus, even SNPs with small effect sizes ($p > 5e-8$, below statistical significance, having small impact on disease susceptibility) can contribute to overall disease phenotypes. Unlike for protein-coding SNPs, there is no uniform way to decode functional impact of non-coding SNPs. Computational approaches elucidate the role of coding variants on human diseases

utilizing well-established properties of protein-coding DNA, and novel approaches need to uncover how previously orphaned non-coding variants can impact disease (Wong et. al 2021). PheWAS (phenome-wide association study) tests for any meaningful genotype-phenotype association between GWAS SNPs and various disease phenotypes (Diogo et al. 2018). It still is challenging to link GWAS or PheWAS non-coding disease-risk SNPs to potential disease genes and understand their downstream functions (e.g. impact on genes, cellular, molecular mechanisms), and association with deeper disease phenotypes like susceptibility and progression.

To address these issues, we performed an integrative analysis (SNPheno pipeline) of multi-omics to reveal genes, functions and GRNs from AD and/or severe Covid-19 GWAS SNPs to AD phenotypes for the three brain regions as mentioned above (**Figure 2.1**, **§2.2**). Given a brain region, we built a gene co-expression network using population gene expression data from an AD cohort and identified co-expressed genes and modules associated with AD phenotypes. We then integrated chromatin interaction data (e.g. High-throughput chromosome conformation capture (Hi-C)) and TF-gene expression relationships to predict TFs regulating co-expressed genes by binding to the regulatory elements that control these genes. Our machine learning (ML) analysis prioritized 36 AD–Covid candidate genes for predicting Covid severity and we evaluated further their ability to predict AD. Finally, we identified risk SNPs altering these TFBSs and analyzed their impact on our GRNs and AD phenotypes. We emphasized subnetworks and regulatory SNPs associated with our predicted AD–Covid genes. Thus, our analysis may provide deeper insights into molecular causes of neuroimmunology pertaining to AD, Covid-19 severity, neuro-Covid and AD–Covid.

§ 2.2 Materials and Methods

2.2.1 SNPheno: Our pipeline of integrative analysis for predicting gene regulatory mechanisms from AD and/or severe Covid-19 risk variants to AD phenotypes

Our analysis can be summarized as a pipeline to predict SNP-effected-GRNs (linking SNPs to GRNs) from disease risk variants to phenotypes (**Figure 2.1**). SNP-effected-GRNs for specific phenotypes link

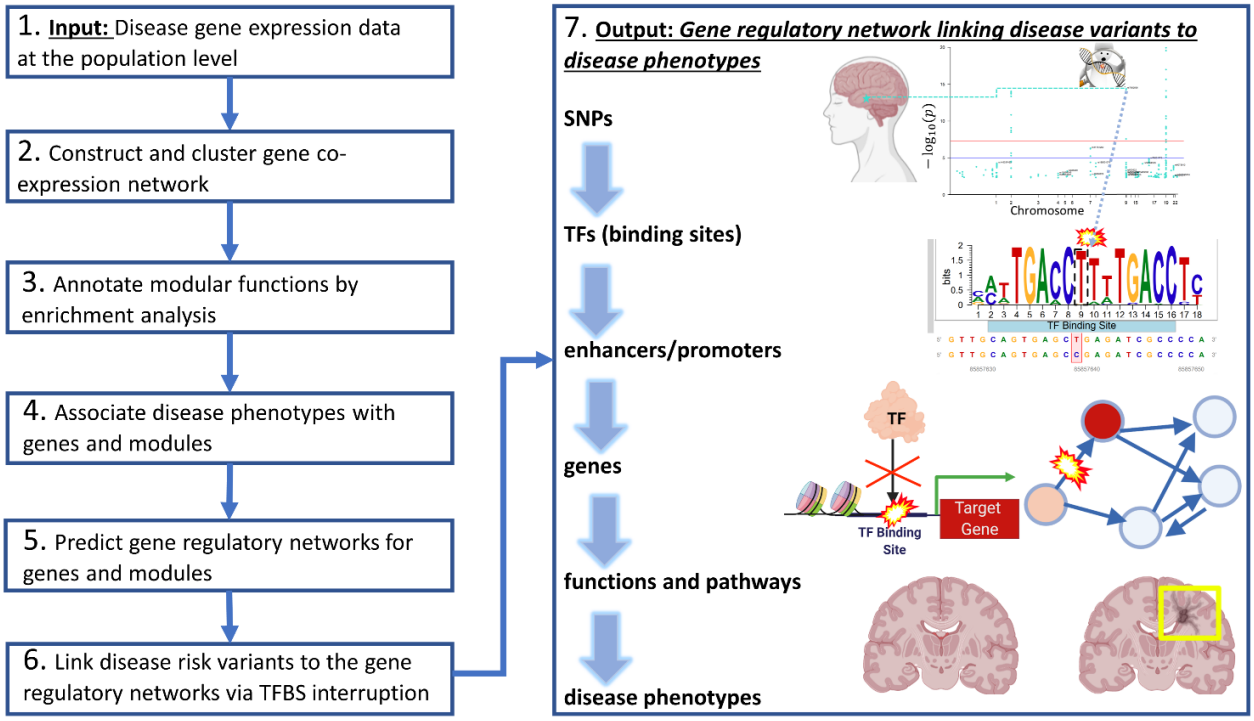
disease risk SNPs, non-coding regulatory elements and TFs to genes and genome functions, providing comprehensive mechanistic insights on gene regulation associated with disease phenotypes. Specifically, the pipeline includes the following steps. Here, our analysis is open-source available at https://github.com/daifengwanglab/ADSNPheno. We use human reference genome: hg19 (GRCh37: Genome Reference Consortium Human Build 37) for genomic coordinates and our following analysis.

- *Step 1:* Input population gene expression data of individuals and clinical information on AD phenotypes such as Braak staging and progression.
- *Step 2:* Input data are used to construct a gene co-expression network linking all possible gene pairs. Network edge weights are correlations of gene expression profiles across input samples. The network is clustered further into gene co-expression modules. Genes in a module are likely to share similar functions and be co-regulated by specific regulatory mechanisms.
- *Step 3:* Annotate gene co-expression modular functions and biological pathways by enrichment analyses of genes in the given module (using various biological resources).
- *Step 4:* Associate modules and genes with AD phenotypes of the input samples, revealing potential driver genes (e.g. hubs) and modules for these phenotypes.
- *Step 5:* Predict gene regulatory networks (GRNs) for genes and gene modules. We apply multiple computational methods to predict GRNs that link TFs to non-coding regulatory elements (e.g. enhancers, promoters) to genes and modules, providing regulatory mechanistic insights on AD genes and modules.
- *Step 6:* Link disease risk variants (e.g. Single Nucleotide Polymorphisms (SNPs)) to the gene regulatory network. Our pipeline identifies functional AD and/or severe Covid risk SNPs that alter (increase or decrease) TF binding to TF binding sites (TFBSs) in regulatory elements in the GRN i.e. regulatory SNPs. We can then connect these non-coding regulatory SNPs to genes and modules and then to AD phenotypes and biological enrichments.
- *Step 7:* Output a SNP-effected-GRN that links AD and/or severe Covid risk SNPs, non-coding regulatory elements, TFs to genes and their gene modules, genome functions (via module enrichment

analysis in *Step 3*) for AD phenotypes in the input data. This network has SNP-Regulatory Element-TF-gene-module-phenotype links.

This SNP-effected-GRN highlights how non-coding disease SNPs alter TF binding on or near TFBSs in regulatory elements for TGs, which belong to certain gene co-expression modules (based on shared expression dynamic patterns) that have various biological functions and have respective associations with various AD phenotypes in the patient population data.

Figure 2.1 – SNPheno: Integrative analyses to predict gene regulatory networks from disease risk variants to phenotypes



Primarily, this analysis consists of 7 major steps as a pipeline: SNPheno.

- Step 1: it inputs the population gene expression data with phenotypic information.
- Step 2: It uses that expression data to construct and cluster gene co-expression networks → gene modules.
- Step 3: It performs enrichment analysis for these gene modules.
- Step 4: It links genes and modules to various phenotypes from the input population.
- Step 5: It predicts the Transcription Factors (TFs) and regulatory elements (e.g. TF binding sites along enhancers and/or promoters) that regulate genes and co-regulate modular genes as a GRN.
- Step 6: It further finds disease risk variants [e.g. Genome-Wide Association Studies (GWAS) Single Nucleotide Polymorphisms (SNPs)] that alter the binding sites of TFs from this GRN.
- Step 7: Finally, it outputs a SNP regulatory network (SNP-effected-GRN) linking functional non-coding disease risk variants to impacted TFs and enhancers/promoters to regulated genes and modules to enriched functions and pathways to disease phenotypes. This network thus provides a deeper understanding of gene regulatory mechanisms in diseases. As a demo, in this paper, we applied this pipeline to Alzheimer's disease (AD) population datasets from different brain regions.

We predicted brain-specific GRNs for various AD phenotypes such as progression stages. Then, we built SNP-effected-GRNs by mapping single nucleotide polymorphisms (SNPs) from several AD GWAS datasets and a GWAS related to Covid-19 severity in Covid-positive individuals to these GRNs.

2.2.2 Population gene expression data and data processing in Alzheimer's disease (AD)

We applied this pipelined analysis to post-mortem human AD population gene expression data for 3 major AD brain regions: Hippocampal CA1 (Hippocampus), LTL, DLPFC. We removed lowly expressed genes (0 variance; relative weights below 0.1), using the goodSamplesGenes() function in the weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath 2008) package in R. There are 12,183 shared genes across these 3 regions, including non-coding genes. We did not pre-adjust gene expression data using covariates (e.g. patient metadata) as those are used downstream as phenotypes. We performed feature engineering to create additional phenotypes for the human samples. We processed the data as follows, striving to meet quality control standards.

Hippocampus: We used microarray gene expression dataset (GSE1297) (Blalock et al. 2004), which had total RNA expression values for 22,283 HG-U133 Affymetrix Human Genome U133 Plus 2.0 Microarray Identifier probes for 31 individual samples (9 control (no AD) and 22 samples in various AD stages: 7 initial, 8 moderate, 7 severe). We used GEOquery (Davis et. al 2007), hgu133a.db (Carlson et. al 2016), hgu133acdf (Project TB, 2015) and Affy (Gautier et. al, 2004) R packages to download raw gene expression data and perform Robust Multichip Average (RMA) normalization (Fan et. al 2013) to account for background and technical variations among the samples. We mapped microarray probes to genes, averaging values that mapped to the same gene Entrez ID and removing unmapped probes. We applied a $log_2(x+1)$ transform to the gene expression data (the x) and then standardized that data by R's (R Core Team, 2021) scale() function. The final Hippocampus expression data has 13,073 genes for the 31 samples.

LTL: We used normalized bulk RNA-Seq dataset (GSE159699) (Nativio et. al 2020) with total RNA expression values for 27,130 different genes for 30 individual samples. This group of individual samples includes 18 control samples (8 young (below age 60), 10 old (above age 60)) and 12 old samples

with advanced AD. We applied a $log_2(x+1)$ transformation to the data. The final LTL gene expression data has 25,292 genes for these 30 samples.

DLPFC: We used FPKM (Fragments per kilobase of exon per million mapped fragments) gene expression data from the ROSMAP Study (synapse.org ID: syn3219045) (Perry et. al, 2018). We found that 638 out of 640 individual RNA-Seq samples have mapped phenotypes. For instance, for the final consensus cognitive diagnosis (cogdx) phenotype on cognitive impairment: we have 201 samples in Group 1 (none), 168 in Groups 2–3 (mild), 269 in Groups 4–6 (AD/other dementia). We applied a $log_2(x+1)$ data transform and then standardized data with R's scale() function (R Core Team, 2021). The final DLPFC gene expression data has 26,014 genes for the 638 samples.

2.2.3 Regulatory elements and Chromatin interactions in human brain regions

Epigenomic data has identified a variety of regulatory elements like enhancers and promoters. Chromatin interaction data (e.g. Hi-C) further revealed interactions among enhancers and gene promoters. Thus, we integrated recently published epigenomic and chromatin interaction data to link enhancers to genes (via promoters). For the Hippocampus, we obtained enhancers and promoters from Brain Open Chromatin Atlas (Fullard et. al 2018) and promoter-based interactions from GSE86189 (Jung et. al 2019). We used R package TxDb.Hsapiens.UCSC.hg19.knownGene (Team BC, Maintainer BP, et. al 2019) to retrieve promoter start and stop positions of genes in the LTL and DLPFC, using a short ultra-conserved promoter length of 5,000 base pairs upstream of the protein-coding start site on the DNA (Rödelsperger et al. 2009). GSE130746 (Nativio et. al 2020) H3K27ac (DNA Histone H3 protein acetylation of the lysine residue that is found at the N-terminal position 27 for H3) data (used for the LTL) has information on the gene, distance from the histone H3K27ac epigenetic mark to that gene's Transcription Start Site (TSS), enhancer start and end positions; our final LTL enhancers were at least 1 kilobase pair (kbp) away from the TSS. We used PsychENCODE (Gandal et. al 2018) enhancers and interacting enhancer-promoter pairs for the DLPFC.

2.2.4 Weighted Gene co-expression network analysis (WGCNA)

We applied WGCNA (Langfelder and Horvath 2008) to population gene expression data to construct and cluster gene co-expression networks into gene co-expression modules (minimum module size = 30 genes; no modules were merged). Then, we applied an additional K-Means clustering step based on code (Botía et. al 2017) and methodology previously utilized and proven to improve conventional WGCNA module assignments and functional enrichments (Lear et al. 2023), in applications like finding brain-specific cell-type marker enrichments. This step utilizes modular eigengenes (MEs) from WGCNA modules as initial centroids, initial WGCNA gene assignments and computable distance between the genes and MEs for K-Means to re-assign genes to optimal modules (retaining the number of modules originally detected by WGCNA) across iterations till convergence. In total, we obtained 30 gene co-expression modules for Hippocampus (13,073 genes), 56 for LTL (25,292 genes), 35 for DLPFC (26,014 genes).

2.2.5 Enrichment analyses of gene co-expression modules

Co-expressed genes in the same module are highly likely to be involved in similar functions and pathways. Enrichment analysis has thus been widely used to identify such functions and pathways in a gene module. *P*-values for enrichments were adjusted using the Benjamini–Hochberg (B-H) correction procedure (for multiple hypothesis testing) and enriched terms with adjust *P*-value < 0.05 were selected. Given a group of genes (e.g. from a module) for each brain region, we performed enrichment analysis using multiple tools and their respective gene databases (e.g. Metascape (Zhou et al. 2019), g:Profiler (Reimand et al. 2007), WGCNA (Langfelder and Horvath 2008), rentrez (Winter et al. 2020), Bader Laboratory (Bader et. al, 2021), Maayan Laboratory (Rouillard et al. 2016), ABAEnrichment (Grote et al. 2016), Psygenet2r (Gutierrez-Sacristan et al. 2023), TissueEnrich (Jain and Tuteja 2019), ClusterProfiler (Wu et al. 2021; Yu et al. 2012), CellMarker (Zhang et al. 2019)). **Table A.1** (SaniyaKhullar 2024) lists the hundreds of data sources used for enrichment analyses. Since we used multiple tools for enrichment analysis, a gene module could have many $-\log_{10}(adjust\ \mathrm{P})$ enrichment values for a given enriched term; in that case, we used the highest enrichment value for that term for the module. To visualize enriched terms for a

phenotype in a brain region, we averaged non-zero $-\log_{10}(adjust\ \mathrm{P})$ values for only the gene modules that are significantly positively correlated (Pearson $r > 0$, $P < 0.05$) with that phenotype.

2.2.6 Association of genes and modules with AD phenotypes

AD patients experience memory loss, cognitive decline, and weak executive function, as reflected in their poor Mini-Mental State Examination (MMSE) results. These changes are brought about by an accumulation of Amyloid plaques between neurons, Neurofibrillary Tangles (NFTs) within neurons, and neuroinflammation that ultimately lead to massive neurodegeneration (Riddle 2012). We further associated genes and modules with these key AD developmental phenotypes, including: AD stages and progression (moderate stage, severe stage, AD Progression), healthy/resilient (Control individuals or resilient individuals with better cognitive abilities despite advanced AD pathology), APOE genotype (E4/E4 is a huge AD risk factor, while E2/E2 is protective (Safieh et. al, 2019)), Braak staging (stages from 1 to 6, with 6 linked to severe neuropathological damage and spread of NFTs across the brain), accumulation of Amyloid plaques between neurons (neuritic plaque accumulation measured by CERAD score), cognitive impairment level. We associated gene co-expression modules with all possible AD phenotypes from the input data, by computing the pairwise correlations of each modular eigengene (ME) with each phenotype. WGCNA's MEs are the first principal components of modular gene expression; an ME is a vector representing gene expression levels of input samples and is the likeliest gene expression pattern of the genes in that module. We used WGCNA's moduleTraitCor() and moduleTraitPvalue() functions to correlate these MEs with phenotypes, finding the most significantly positively associated phenotypes for our gene modules for our analysis (P-value < 0.05, positive correlation r). Our modules of interest i.e. 'phenotype-enriched modules' are positively correlated with at least one AD-related phenotype (including the control stage phenotype since such modules may typically be down-regulated in expression during AD progression). We performed similar analysis (that we used for the MEs) for each of our genes using the expression data for that given gene to find significant phenotypes positively associated with that gene in that brain region. We used gene co-expression networks to examine the

relationship between genes and AD phenotypes and identify potential driver (hub) genes for modules (based on the degree of connectivity for each gene in its respective gene co-expression module).

2.2.7 Prediction of gene regulatory networks (GRNs) from multi-omics

GRNs, a key molecular mechanism, fundamentally control gene transcription and expression. Co-expressed genes are likely co-regulated by similar GRNs. In fact, studies (e.g. (Allocco et al. 2004)) have observed that a given pair of genes with very strongly correlated gene expression profiles are very likely to have a common TF that binds to their respective promoter regions on the DNA; moreover, higher co-expression correlations among these genes are associated with an increased number of common TFs (Gu et al. 2011). Thus, our analysis integrates multiple methods to predict GRNs from gene expression data. We predicted GRNs in brain regions using not only gene expression data but also chromatin interaction data to link TFs to regulatory elements to target genes (TGs)/modules. For our full DLPFC GRN, we used the published PsychENCODE GRN (Elastic Net regression weight cutoff: 0.1) filtered for genes in the DLPFC gene expression data (Wang et al. 2018a). Our full GRNs linked TFs to regulatory elements (REs) (enhancers/promoters; chromosome #: regulatory region start—end) to TGs.

We used these 4 steps to construct our full Hippocampus GRN and full LTL GRN. First, we identified REs that potentially interact using recent chromatin interaction data (Hi-C) and Step 1 of the scGRNom pipeline (Jin et al. 2021). Second, we infer TFBSs on the basis of consensus binding site sequences on interacting enhancers and promoters by TFBSTools (Tan and Lenhard 2016) and motifmatchr (Schep and University 2023) using Step 2 of the scGRNom pipeline. We generate a chromatin interaction-based reference network linking TFs to REs (by TFBSs) to TGs (by interactions). Third, using gene expression data for a given brain region, we predicted all possible TF–TG pairs (or TF-modules) with strong expression relationships by applying three widely used tools: RTN (Groeneveld et al. 2023), TreNA Ensemble Solver (Arment S et al. 2021), GENIE3 (Huynh-Thu et. al, 2010) (and TF-gene-module pairs by RTN). Thus, we created a gene expression-based network by combining TF–TG pairs found by ≥ 2 of these 3 tools. Fourth and finally, we mapped the gene expression-based network TF–TG pairs to the TF-

TG pairs in the chromatin interaction-based reference network. The full GRN (for Hippocampus or LTL) thus contains TF–TG pairs found in both the chromatin interaction and the gene expression-based networks (**§A.1** (SaniyaKhullar 2024) has more details).

For each of the 3 brain regions, we built final GRNs by using our prior analysis (see earlier), which had assigned TGs to gene co-expression modules and associated the modules with AD phenotypes and biological enrichments. This prior analysis provided richer annotations for TGs in our full GRNs. Our final GRN for each brain region comprehensively linked TFs to non-coding regulatory elements to TGs and these TGs to gene modules to AD phenotypes/enrichments.

2.2.8 Identifying AD-Covid GRNs and genes using GRNs and gene modules

To investigate potential mechanistic interplays between AD and Covid-19, we compared AD (hsa05010) and Covid-19 (hsa05171) KEGG networks (Kanehisa and Goto 2000) and found AD–Covid mechanisms like: NF-κB, Inhibitor of Nuclear Factor Kappa B Kinase (IKK), c-Jun N-terminal Kinase (JNK), Interleukin-6 (IL-6), Phosphoinositide 3-Kinase (PI3K), Tumor Necrosis Factor alpha (TNFa), TNF Receptor (TNFR). We found a statistically significant overlap of 22 genes between both KEGG networks based on a hypergeometric test (7,559 human genes in KEGG universe, 384 human KEGG genes in AD, 232 human KEGG genes in Covid-19). The 22 shared KEGG genes correlate highly with AD phenotypes in different brain regions. This motivated us to find neuroimmunology genes in AD–Covid. We used Pathview (Luo and Brouwer 2013) to visualize correlations of KEGG network mechanisms with AD phenotypes. For each region, we constructed an AD–Covid gene list using its respective final GRN and gene co-expression modules as follows. First, we built an AD–Covid GRN: a subnetwork of GRN with TFs regulating and/or TGs of the 22 shared KEGG genes such that each GRN edge contains at least 1 shared KEGG gene. 2nd, we filtered these AD–Covid GRNs to only include genes that belong to an 'AD-phenotype enriched' gene module. Hence, genes in our AD–Covid GRN were either 1 of the 22 shared KEGG genes or directly linked to them by a GRN link. Moreover, these AD–Covid genes had altered expression dynamic patterns associated with AD. Thus, we built four AD–Covid gene lists: LTL, DLPFC, Hippocampus, combined list of the 3. These 4 lists were later used to predict Covid-19 severity (see next).

2.2.9 Gene expression analysis and machine learning (ML) prediction for Covid-19 severity from AD-Covid gene regulatory networks

To gauge the clinical performance of our AD–Covid genes in terms of predicting Covid severity (proxy for immune system dysregulation), we looked at recent population RNA-seq gene expression data of human Covid-19 blood samples (GSE157103) (Overmyer et. al, 2021). We median normalized this data (19,472 genes) and applied differential expression analysis by DESeq2 (Love et al. 2023) between 50 severe (Intensive Care Unit (ICU)) and 50 non-severe (non-ICU) Covid patients. Aside from applying differential expression analysis to find individual-associated differentially-expressed genes (DEGs), we performed machine learning (ML) analysis to determine if any of our four AD–Covid gene lists (from our AD–Covid GRNs) and the respective normalized blood gene expression data could predict the probability of severe Covid (being in the ICU) for Covid patients better than a benchmark list of Covid genes could. We used a SVM classifier model (linear kernel, balanced class weights, on the basis of Python's Scikit-Learn (Pedregosa et al. 2011) svm.SVC package) to output the predicted probabilities of severe Covid for Covid samples. We randomly partitioned our data using an 80–20 training–testing split with 80 samples (40 ICU, 40 non-ICU) in training data and held out 20 samples (10 ICU, 10 non-ICU) in test data. Stratified 5-Fold Cross Validation (CV) was used to calculate training classification accuracy; each fold held out 16 samples (8 from each class) for validation and trained an SVM model on the remaining 64 samples (32 from each class). Input data used to build each model was the median normalized Covid gene expression data for the respective selected genes (features) for the training samples. We did not use age and gender as predictors given their low correlations with Covid severity.

For our ML analysis, we gathered a benchmark list of 18 known and published Covid genes from 4 studies (Hu et. al, 2021; Pairo-Castineira et. al 2020; Hou et al. 2020; Kong et al. 2020). A study (Hu et. al, 2021) used U.K. Biobank GWAS and Covid mortality data to discover 8 genes associated with high Covid mortality: *DNAH7, CLUAP1, DES, SPEG, STXBP5, PCDH15, TOMM7, WSB1*. Another study (Pairo-Castineira et. al 2020) has identified 7 risk genes (*OAS1, OAS2, OAS3, TYK2, DPP9, IFNAR2, CCR2*)

associated with life-threatening Covid outcomes (e.g. inflammatory organ damage). Numerous studies (Hou et al. 2020) implicate SNPs in *ACE2* and *TMPRSS2* genes as risk factors for Covid susceptibility. Another Covid severity study used a Random Forests ML model and has identified *VEGF-D* as the most predictive indicator (Kong et al. 2020). To build our benchmark model, we first performed RFE CV (RFECV) on a SVM model using these 18 benchmark genes to calculate the accuracy of adding a gene to the model and optimal number of genes to use i.e. smallest number of genes with the maximum stratified 5-fold CV training accuracy when classifying ICU versus non-ICU Covid patients. Second, we ran RFE on a SVM model with that optimal number of genes to select the predictive genes from the training data. Third, we used these selected benchmark genes to train another SVM model as our benchmark model. We fixed all models to use this same number of genes to help facilitate direct comparison of the predictive models. We performed the second and third steps instead on each of our respective input AD–Covid gene lists to build our four AD–Covid models. Thus, we built 5 models to predict Covid severity: benchmark, combined, Hippocampus, LTL, and DLPFC. We compared the prediction performance of each of our four AD–Covid models with that of the benchmark model using: accuracy, AUC and Decision Curve Analysis (DCA, **§A.1** (SaniyaKhullar 2024)) on training and test (generalize potential clinical impact of models) data. For each model, we report training metrics by averaging values across all five stratified folds. We flagged 'AD–Covid genes' used in any of our four AD–Covid models (predictive for severe Covid) as potential candidate biomarkers for AD–Covid-neuroimmunology.

2.2.10 ML prediction for AD & Covid severity from AD-Covid genes

We analyzed the performance of our AD–Covid genes (those from among our 4 AD–Covid models) for predicting AD on a new human population cohort (GSE125050) (Srinivasan et al. 2020) of 22 AD and 21 control postmortem Superior Frontal Gyrus (SFG) tissues in the frontal cortex (linked with AD pathology). That study isolated RNA-seq data for 4 brain cell types (neurons, astrocytes, endothelial cells, microglia) from SFG tissues. We pooled raw gene expression data for these 4 cell-types (62 control, 46 AD samples, **Table A.2** (SaniyaKhullar 2024)) for our task. For each cell type, we held out three AD and three Control cell-type samples for testing (total: 12 AD, 12 Control). The remaining 84 samples were

used to train a model (Python's Scikit-Learn (Pedregosa et al. 2011)) Logistic Regression (LR) package, liblinear solver, balanced class weights) to predict AD or control for a given sample. Our features were pooled gene expression data values for AD–Covid genes and four dummy (0 or 1) features noting the cell-type for each sample. We built another LR model as a benchmark, only changing the gene features we used, which were now 597 AMP-AD (AMP-AD (Agora)) nominated genes for AD identified in the SFG gene expression data. The AMP-AD consortium flagged these AMP-AD genes as good targets for AD treatment and/or prevention based on computational analyses in previous studies of human samples. We kept shared and common AMP-AD and AD-Covid genes as gene features to train both LR models. We compared the test performance of the optimal AD–Covid LR model and benchmark AMP-AD LR model to better quantify the effectiveness of our AD–Covid genes in predicting AD as well. Furthermore, we noted our AD–Covid genes that were DEGs in recent single-cell transcriptomic data (Mathys et. al 2019) analysis for AD pathology versus controls in Excitatory (ExNs) and/or Inhibitory (InNs) neurons.

2.2.11 Linking Genome-Wide Association Study (GWAS) SNPs for AD and for Covid-19 severity to gene regulatory elements

GWAS have identified genetic risk variants associated with diseases like AD. However, most AD SNPs lie on non-coding regions, hindering finding AD genes and understanding downstream disease functions. We consider SNPs with $P < 5e\text{-}5$ to include candidate disease SNPs via interrupting gene regulation at large (**Table A.3** (SaniyaKhullar 2024)). We looked at summary statistics of 26,969 AD risk GWAS SNPs across 5 studies (Jansen et. al 2019; Kunkle et. al 2019; Wightman et. al, 2021; Turley et al; Bellenguez et. al 2022) and 1,642 SNPs from the 7th round of GWAS meta-analyses related to severity across all Covid-19 positive human populations (COVID-19 Host Genetics Initiative 2022): 16,512 hospitalized cases (severe) versus 71,321 not hospitalized controls (non-severe). Risk SNPs for a condition have a positive effect size in the GWAS, i.e. the SNP is associated with higher disease phenotypes or traits. On the other hand, protective SNPs have a negative effect size, so the SNP is associated with decreased disease phenotypes or traits.

We mapped SNPs to regulatory elements (REs) in the GRNs via altered TFBSs (**Figure 2.1**, pipeline step 6). We overlapped 28,597 AD and/or severe Covid risk SNPs (14 common) with regulatory elements (enhancers, promoters) in our final GRNs. MotifbreakR (Coetzee et al. 2015) identified 24,576 SNPs altering TFBSs of 791 TFs. These regulatory SNPs either increase TF affinity for the TFBS (based on TF sequence-specific motifs) or interrupt and subsequently decrease TF binding to that RE. We linked these SNPs to TGs from REs with altered TFBSs, adding a 10 kilobase (kbp) and 2 kbp buffer extension to the start and end positions of enhancers and promoters, respectively. Thus, we mapped our SNPs to our final GRNs. Our SNP regulatory network (SNP-effected-GRN: SNP effect on our final GRN) comprised our predicted SNP–RE–TF–TG–Module–Phenotype links. We used expression quantitative trait loci (eQTL) data (associates SNP with changes in TG expression (Liu et. al 2022)) from various sources (**Table A.4** (SaniyaKhullar 2024); tissues: brain (PsychENCODE Consortium; THE GTEX CONSORTIUM 2020; Patel et. al 2021)/blood(Patel et. al 2021); cell-types: brain (Byrois et. al 2022; Zeng et al. 2022)) to annotate SNP-effected-GRN links with this external SNP-TG validation as highly-confident; our SNP-effected-GRN may explain GRN mechanisms behind these causal eQTL links. An eQTL is a genetic region that may help explain how SNPs impact variety in gene expression phenotype levels for local (cis) or distal (trans) genes (Nica and Dermitzakis 2013; Liu et. al 2022). Further, we performed linkage disequilibrium (LD) via LDlink (Machiela et. al 2015)) (GRCh37 genome, all human populations) to correlate a pair of SNPs (on the same chromosome); linked SNPs have significantly correlated alleles and tend to be non-randomly inherited together in all populations. **§A.1** has more details and a framework for analyzing our SNP-effected-GRN.

§ 2.3 Results

2.3.1 Gene co-expression network analysis reveals gene expression dynamics for AD phenotypes across multiple brain regions

First, we applied our analysis to population gene expression datasets of three major brain regions relating to AD: Hippocampal CA1 (Hippocampus), LTL and DLPFC (**§2.2: Materials and Methods**). We identified several gene co-expression modules showing specific gene expression dynamic changes for

various AD phenotypes (**File A1**: Hippocampus, **File A2**: LTL, **File A3**: DLPFC (SaniyaKhullar 2024)), implying potential underlying gene regulatory mechanisms associated with the phenotypes. Given a brain region, we constructed and clustered a gene co-expression network to a set of gene co-expression modules. In a gene co-expression network for a brain region, nodes (or vertices) are genes and each edge represents that two respective genes have correlated gene expression profiles across the samples (i.e. co-expression). There are likely groups of co-expressed genes within the network that form densely connected sub-networks (gene modules). Genes in a module share similar gene expression dynamics in that respective brain region for the observed AD phenotypes. Modular eigengenes (MEs) represent expression dynamics for a gene module, using the first principal components of module gene expression matrices.

Hippocampus: 21 of 30 gene modules (9,525 genes) are 'phenotype-enriched' as they are significantly positively associated with at least one key AD-related phenotype. Their MEs show specific expression dynamics (**Figure 2.2A**: 7 select modules; **Figure A.1A** (SaniyaKhullar 2024): all 30 modules). Pink and lightyellow modules have high gene expression values for Controls and cluster together. On the other hand, greenyellow, yellow, tan and magenta modules cluster together given their high expression in AD. Next, we used expression dynamic patterns to link modules to phenotypes (**Figure 2.2B**: 7 select modules; **Figure A.1B** (SaniyaKhullar 2024): all 30 modules) by significant positive correlations. The tan module has the highest severe AD correlation ($r = 0.68$). The midnightblue module is significant for Braak 4 stage (mild dementia), the lightyellow module for cognitive resilience. The greenyellow module significantly correlates with AD, AD progression, moderate/severe AD, cognitive impairment, Braak 6 stage (severe dementia).

LTL: 28 of 56 co-expression modules are phenotype enriched. We highlighted five MEs in **Figure 2.2C (Figure A.1C** (SaniyaKhullar 2024): all 56 modules). The sienna3 module has higher expression values for old and young Controls. Orange, magenta, and yellow modules cluster together with high expression in AD. As shown in **Figure 2.2D** for the same 5 select modules (**Figure A.1D** (SaniyaKhullar 2024): all 56 modules), the sienna3 module correlates positively with Controls ($r = 0.63$)

and being asymptomatic for dementia or any other AD-related symptoms ($r = 0.55$). Yellow, orange and magenta modules associate with aging, AD/Braak progression, neuritic plaques; the orange module has $r = 0.72$ for AD/dementia.

DLPFC: The sample size in the DLPFC, which is 20-fold larger relative to those of the other 2 regions, likely attributes to the comparatively lower module-phenotype correlations we observe in the DLPFC. Still, we see significantly correlated modules with select AD phenotypes and highlight 6 of 35 modules (all 35 are phenotype-enriched) (**Figures 2.2E-F**, $P < 0.05$). The tan module is associated with the worst APOE genotype (E4/E4) and with age for diagnosis of AD; royalblue and green DLPFC modules correlate with severe AD based on last MMSE score. In terms of better and healthier outcomes, the darkolivegreen module is significant for Controls, higher MMSE scores, cognitive resilience. **Figures A.1E-F** (SaniyaKhullar 2024) show results for all 35 gene co-expression modules in the DLPFC.

Our gene modules across regions uncover gene expression dynamic patterns across phenotypes, suggesting that genes in a module are likely involved in similar functions and pathways. To understand this, we performed module enrichment analysis as follows.

2.3.2 Eigengenes and enrichments of co-expression modules reveal hub genes, gene functions, and pathways in AD phenotypes

We performed gene set enrichment analyses (**§2.2: Materials and Methods**) to understand better the biological functions, diseases, pathways, structures and other observed phenomena of our modules and link them to various AD phenotypes (**Figure 2.2**). Healthy phenotypes are Control, cognitive resilience, protective APOE E2/E2 genotype. Our brain region module enrichments underscore the role of the immune system and neuroimmunology among other factors in AD progression and verify that the phenotype correlations we detected for our gene modules may indicate true biological signals. **Figure A.2A-C** (SaniyaKhullar 2024) shows enrichment results for select gene co-expression modules, for each of the 3 brain regions, which have strong correlations r with relevant phenotypes in the population.

In AD, the Hippocampus (**Figure A.3A**, **File A1** (SaniyaKhullar 2024)) has a major loss in volume, neurogenesis, memory, neuron density (Bartsch et. al, 2011). Healthy gene modules are enriched with

synaptic plasticity, dendrite development, calcium signaling. Perhaps resilient individuals are protected from microsatellite instability and amyloid accumulation. Age and AD progression modules are associated with abnormal innate immunity, Covid-19 spike protein, NF-κB pathway (overexpressed in AD (Jha et. al 2019)), activation of JNK and MAPK cascade [active in AD, involved in tau phosphorylation, neuroinflammation (Lee and Kim 2017), synapse dysfunction, neuron death (Lu et. al 2014). Severe AD modules are associated with metabolic processes (Mapstone et al. 2020), immune memory, interferon signaling [high in AD mouse Hippocampus (Naughton et al. 2020); this response to amyloid may activate microglia, initiating neuroinflammation and synapse loss in neurons (Roy et. al 2020).

The LTL (**Figure A.3B**, **File A2** (SaniyaKhullar 2024)) is impacted early in AD (Nativio et. al 2020). Control modules are enriched with Wnt signaling (inhibits tau protein hyperphosphorylation and production of amyloid-beta (Aβ) plaques (Inestrosa and Varela-Nallar 2014)), whose dysregulation may lead to neurodegeneration. In AD and plaque modules, Nlp protein loss from Mitotic centrosomes is enriched (may cause microtubule instability, abnormal cell morphology, AD (Dubey et al. 2015)). We found cell-type and other pathway enrichments in AD progression phenotypes: NF-κB activation, astrocyte projection (glial cell type that is increasingly active near Aβ plaques in AD (Vasile et. al, 2017)), prion pathway (disruption may lead to Aβ plaques (Kellett and Hooper 2009)). Dramatic Histone H4 acetylation epigenetic losses on DNA regions near genes may decrease memory formation during aging and AD in the LTL (Nativio et. al 2020).

The DLPFC (**Figure A.3C**, **File A3** (SaniyaKhullar 2024)) works with the Hippocampus to mediate complex cognitive functions (Diaz Brinton 2012) and has plasticity deficits in AD (Kumar et. al 2017). Microglia exclusively express AD genes like APOE (Hemonnot et. al 2019). Here, APOE2 modules are associated with mitochondrial inheritance ($P < 1e\text{-}16$; whose dysfunction is associated with various brain-related disorders (Shen et al. 2023)) and are shielded from neurotoxins, whereas APOE4 modules are enriched with Aβ response (may regulate microglia (Fullard et. al 2018)), cognitive dysfunction. Promising associations (some with $P < 1e\text{-}58$) for APOE4 and AD-related modules support a crucial role of reactive microglia for AD disease progression. In AD, microglia may change shape, are more phagocytic, go awry

and release pro-inflammatory cytokines, leading to Aβ and neurofibrillary tangles (NFTs) (Kinney et al. 2018b), synapse decline, neuroinflammation, cell death, neurodegeneration (Hemonnot et. al 2019). Our results may shed light on links between APOE4 and neuroinflammation, with enrichments such as: autoimmune diseases (e.g. Wegener's Granulomatosis), synapse pruning, astrocyte activation, microglia, abnormal innate immunity and cytokine levels (DLPFC in AD patients typically has more pro-inflammatory cytokines like IL-1B, linked to Aβ plaques (Kinney et al. 2018b)). Healthy modules are enriched with Electron Transport Chain (altered in AD (Ebanks et al. 2020)), neuron recognition, synapse plasticity, calcium ion regulated exocytosis. Finally, we compared 3 brain regions (**Figure 2.2G**, **Figure A.3D** (SaniyaKhullar 2024)): Braak stage modules are enriched with Ki-1 antigen (tumor marker of activated immune cells regulating NF-κB and apoptosis), focal adhesion (plaques), VEGFA-VEGFR2 (altered levels in AD may impact microglia/neuron survival (Cho et. al 2017)). Control and AD DLPFC/Hippocampus modules share neuroimmunomodulation. AD and Braak stage modules are enriched with blood–brain barrier (BBB), virus attachment, complement system (CS) activation (innate immune-mediated defense altered in AD (Carpanini et. al 2019)), oligodendrocyte differentiation [this change in this glial cell type is linked to neurodegeneration, Aβ accumulation (Quintela-López et al. 2019)].

Figure 2.2 – Gene co-expression modules significantly associated with AD phenotypes show specific expression dynamic patterns across phenotypes and enriched functions and pathways.
Corresponding heatmaps for all modules in the 3 brain regions are in **Figure A.1A-F** (SaniyaKhullar 2024)).

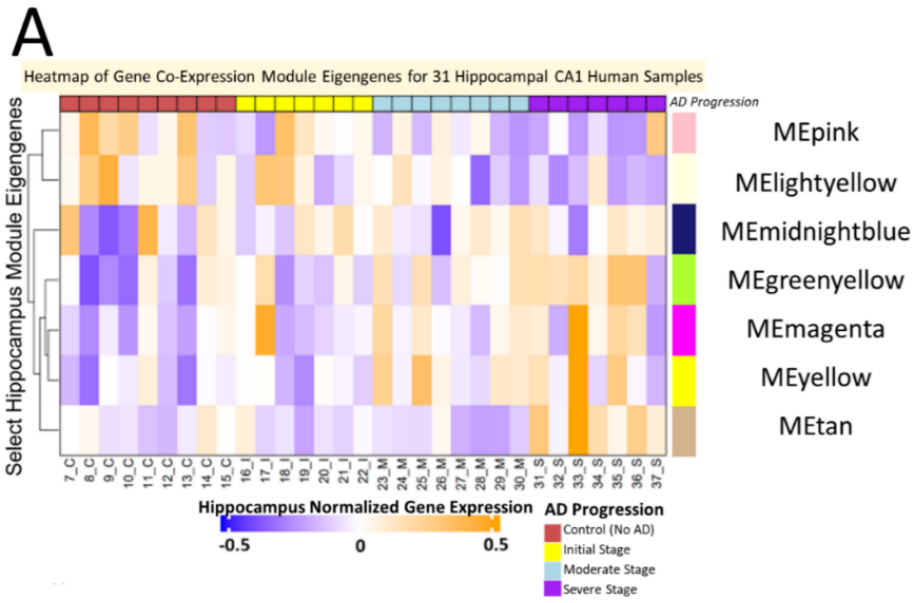


Figure 2.2A) - Module eigengenes (MEs) of 7 gene co-expression modules in the Hippocampal CA1 region where rows: modules and columns: individual human samples. Red: high expression level. Blue: low expression level. On the left hand side of this heatmap is a dendrogram tree based on agglomerative hierarchical clustering so that similar modules (in terms of values for MEs) cluster close together.

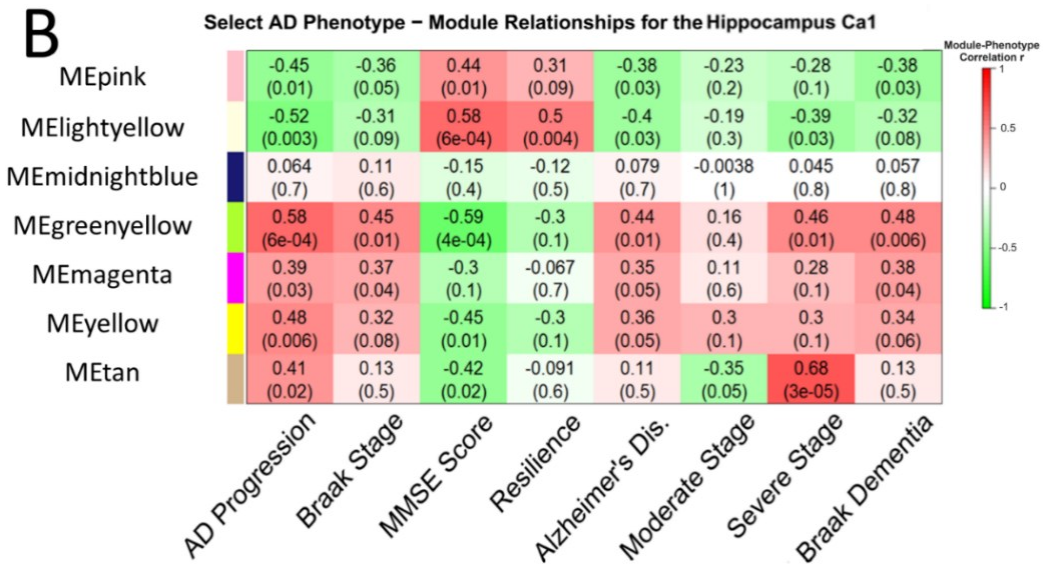


Figure 2.2B) - Shows the correlation coefficients and *P*-values for the same 7 Hippocampal CA1 gene modules and various select AD phenotypes.

File A1 (SaniyaKhullar 2024) contains additional phenotypes. Row: modules. Columns: AD phenotypes. Red: highly positive Pearson correlation ($r > 0$). Green: highly negative correlation ($r < 0$).

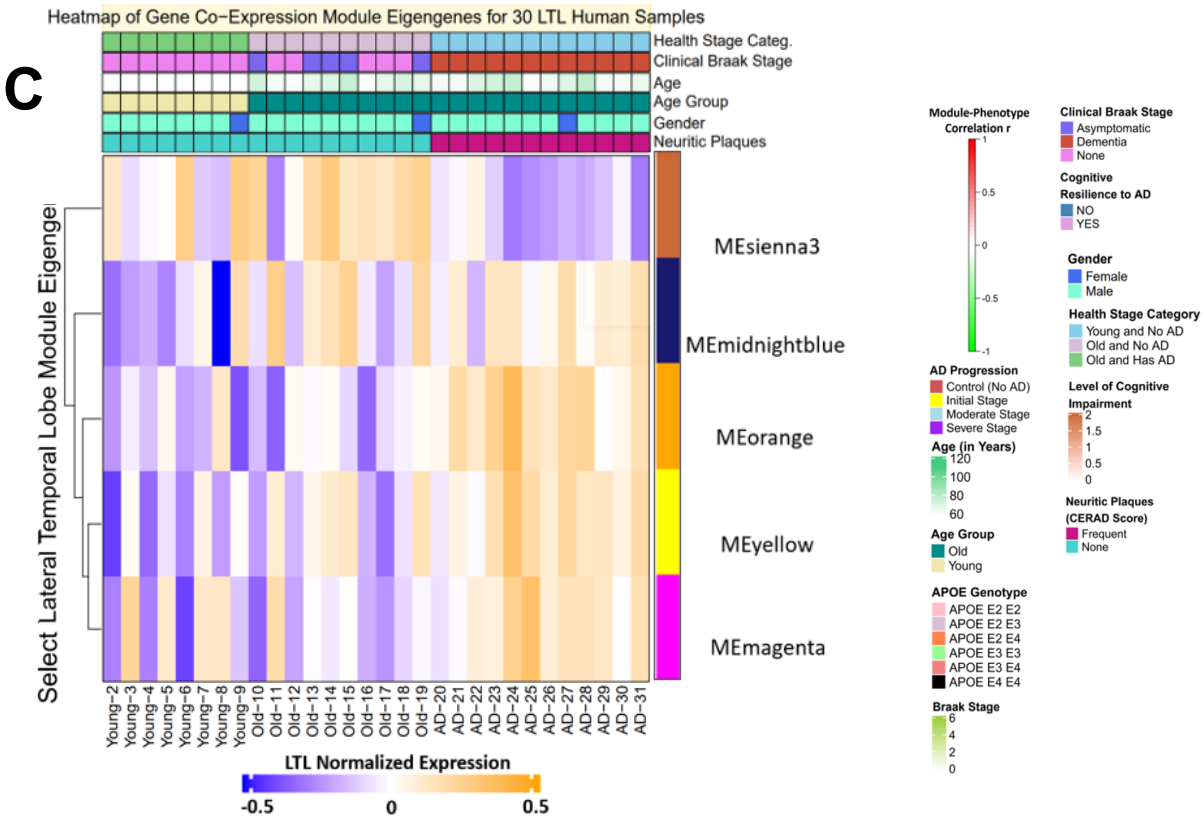


Figure 2.2C) - Module eigengenes (MEs) of select gene co-expression modules in the LTL region. Red: high expression; blue: low expression level. Heatmap for select gene co-expression modules in LTL.

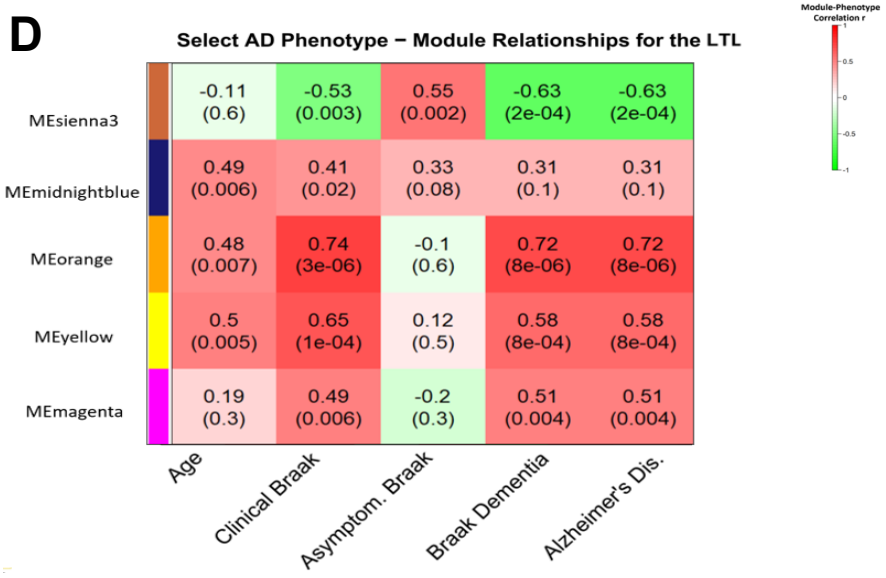


Figure 2.2D) - Correlation coefficients and p-values between for the same modules (for that given brain region as in **Figure 2.2C**) and AD phenotypes in the LTL region.

File A2 (SaniyaKhullar 2024) contains additional phenotypes. Row: modules. Columns: AD phenotypes. Red: highly positive correlation. Green: highly negative correlation.

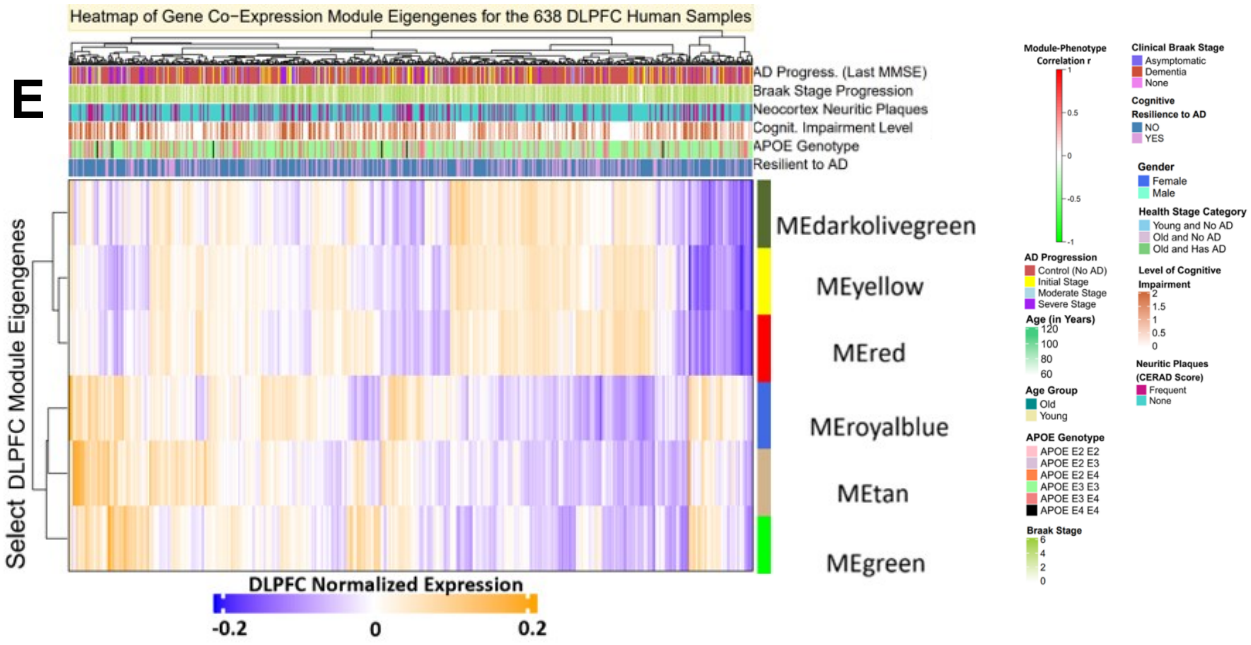


Figure 2.2E) - Module eigengenes (MEs) of select gene co-expression modules in the DLPFC region.

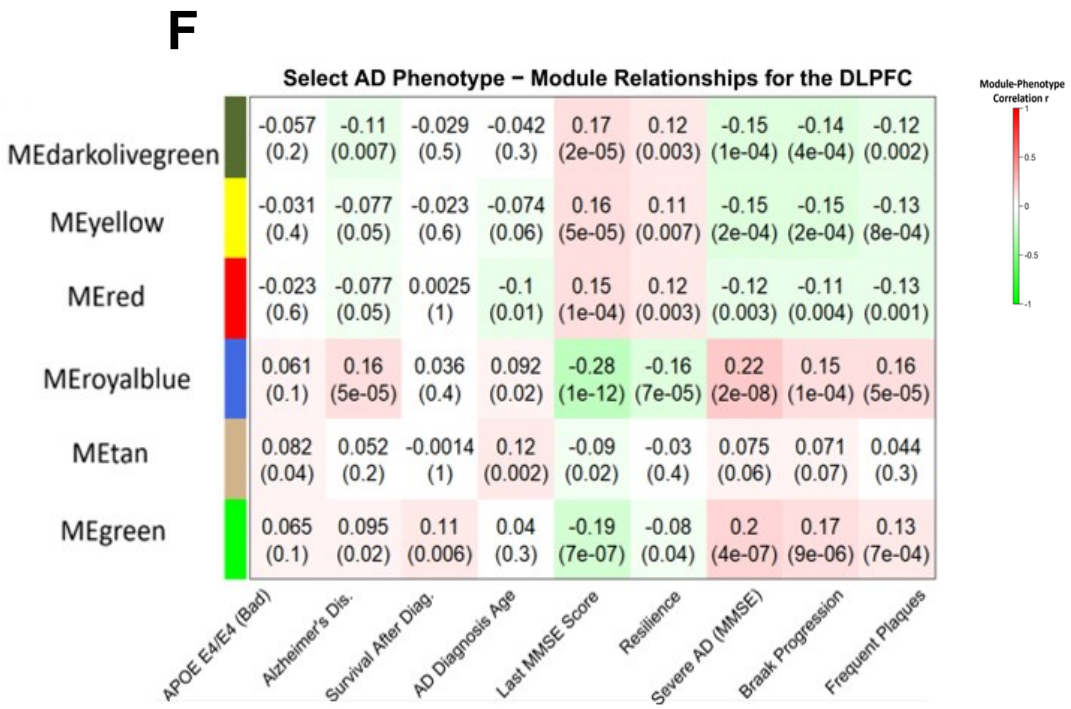


Figure 2.2F) - Correlation coefficients r and p-values between for the same modules (in **Figure 2.2E**) and AD phenotypes in the DLPFC region. Row: modules. Columns: AD phenotypes. **File A3** (SaniyaKhullar 2024) contains additional phenotypes.

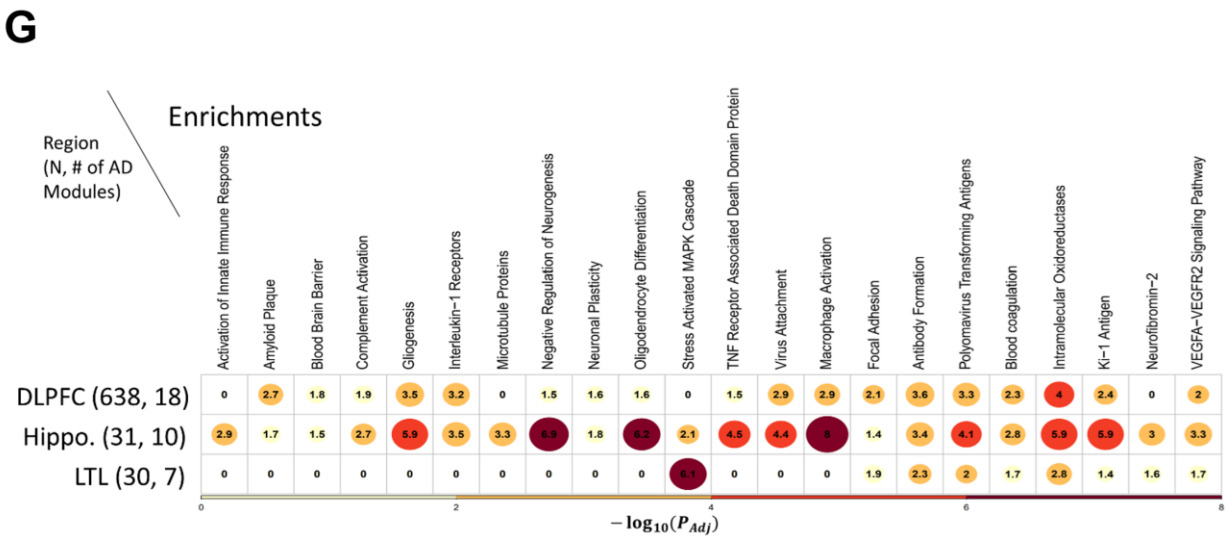


Figure 2.2G) - Shows select biological functions and pathways that are enriched for modules positively correlated ($P < 0.05$, r > 0) with AD across the *3* brain regions.
Values in the cells, circle sizes and gradient color (yellow to red) correspond to the highest enrichment $-\log_{10}(\text{adjust P}-\text{value})$ value of any gene module associated with the phenotype of AD diagnosis (i.e. AD phenotype) from that given brain region. There are $\text{N} = 31$ post-mortem human samples in the Hippocampus, $\text{N} = 30$ in the LTL and $\text{N} = 638$ in the DLPFC.

2.3.3 Prediction of brain-region gene regulatory networks for AD phenotypes

To understand underlying molecular mechanisms regulating gene expression associated with various AD phenotypes, we predicted the GRNs for target genes (TGs) and gene modules of brain regions, especially using multi-omics data (**§2.2 Materials and Methods**). Brain region GRNs link TFs and regulatory elements (e.g. enhancers or promoters) to TGs and co-expressed genes (e.g. from the same gene module). GRN edges can be activation or repression of TGs by TFs, which follow-ups can investigate. These GRNs can be further linked to AD phenotypes significantly associated with TGs and modules. We applied many popular approaches and public databases to predict networks and used their shared predictions to build our highly confident GRNs. We found: 1,043 candidate TFs in the Hippocampus, 1,580 in the LTL, (and 1,588 in the DLPFC), which we input into RTN, GENIE3 and TReNA Ensemble Solver for the Hippocampus and LTL, respectively. **Table A.5** (SaniyaKhullar 2024) shows statistics of

TF–Regulatory Element–TG network nodes and edges. **Files A4** and **A5** (SaniyaKhullar 2024) contain our detailed final Hippocampus and LTL GRN edge lists, respectively.

We found TFs statistically significantly regulate 21 LTL gene co-expression modules and 21 Hippocampus gene modules (**Figure A.4A-D** (SaniyaKhullar 2024)); for example, in the Hippocampus, neurogenesis TF REST regulates a module of 883 genes, NFKB1 regulates one Control module, RELA regulates three gene modules (2 Control modules, 1 AD progression module). REST is induced by Wnt signaling, protects neurons from Aβ-protein toxicity (Lu et al. 2014). During AD, overexpression of AD risk genes may be partly explained by REST's inability to bind to chromatin and repress its target AD genes (Maezawa et al. 2012); this may lead to autoinflammation, immune disorders (Magno et al. 2019a).We also find that in the Hippocampus, ZNF226 regulates 3 modules (2 which are associated with the Control stage) and GATA 2 regulates 4 modules (1 Control Stage module and 2 modules associated with worsening AD phenotypes).

2.3.4 Gene regulatory networks and AD phenotypes associated with shared AD-Covid pathways

Rogue immune responses characterize AD and Covid-19. Microglia, the brain's resident immune cells, become chronically activated in both AD and COVID-19. This chronic activation leads to sustained inflammation within the brain, causing neurocognitive symptoms and neuronal damage. Our hypergeometric test of overlap between both AD and Covid-19 (SARS-CoV-2) Kyoto Encyclopedia of Genes and Genomes (KEGG) networks ($P = 0.0034$) was significant, suggesting that the shared AD–Covid mechanisms are important. COVID-19 disrupts the blood-brain barrier (BBB), a feature also observed in AD. By studying the inflammatory and immune responses triggered by SARS-CoV-2, we can draw parallels to the mechanisms driving AD.

We thus analyzed these shared mechanisms implicated in adverse effects and inflammation in both diseases (Jha et. al 2019), like the NF-κB pathway. The NFKB pathway is a key regulator of immune and inflammatory responses. In mammals, the NF-κB TF family has 5 TFs: NFKB1 (or p105/p50

protein), NFKB2 (or p100/52 protein), REL (or c-Rel), RELA (or p65 protein) and RELB (proto-oncogene near APOE). Reactive Oxygen Species (ROS) activate RELA and NFKB1 TFs. Both TFs then transcribe pro-inflammatory cytokines (e.g. Interleukin-6 (IL-6), IL-1B, TNF), reducing long-term potentiation (LTP) during AD (typically resulting in reduced strength of synaptic signal transmission between neurons, lower synaptic plasticity, memory loss and learning delays) and leading to exaggerated and potentially lethal immune responses in Covid (e.g. tissue injury, hypoxia (Erausquin et. al 2021), hyperinflammation, Acute Respiratory Distress Syndrome (ARDS) (Kircheis et al. 2020; Khullar et al. 2020)) (**Figure A.5** (SaniyaKhullar 2024)). We found that gene expression levels of NF-κB TFs correlate positively with AD severity but negatively with controls in all three regions (Hippocampus: **Figure A.6A** (SaniyaKhullar 2024)). NFKB1 and RELB correlate negatively with controls in three regions, as do NFKB2 and RELA in the DLPFC and Hippocampus (**Figure A.6B** (SaniyaKhullar 2024)). All 5 TFs correlate positively with severe AD in the Hippocampus and two TFs correlate positively with AD in DLPFC (**Figure A.6C** (SaniyaKhullar 2024)). Upregulation of NF-κB TFs may be a key AD–Covid interplay as activation of these TFs is linked to greater inflammation in Covid and in AD (Kircheis et al. 2020). NFKB1 and RELA's severe AD Hippocampus module has immune enrichments like PID (Pathway Interaction Database) IL-1 pathway, abnormal innate immunity, immunoglobulin level. We investigated our GRNs involving NF-κB TFs. **Figure 2.3A** shows shared target genes (TGs) for NFKB1 and/or RELA in the DLPFC and Hippocampus; seven TGs are regulated by both TFs in both regions, like *ANP32B* and *EMP3*. **Figure 2.3B** shows how RELA and NFKB1 indirectly regulate *IL-1B* in the LTL via TFs: TCF3, RFX3, RREB1, IRF1, TP53.

We looked at the SARS-CoV-2 (Covid-19 KEGG: hsa05171) network to analyze how the NF-κB pathway and regulated cytokines may be associated with AD–Covid links and neuroinflammation (**Figure 2.3C**). During Covid-19 infection, the SARS-CoV-2 Spike protein is primed by TMPRSS2, binds to the ACE2 [high expression in brain/macrophages (Kwee and Kwee 2020)] receptor and interacts with AT1R (Angiotensin II Receptor Type 1) to enter and infect the cell (Qiao et al. 2021). Neurons may be directly invaded by SARS-CoV-2 or by systemic infection compromising the blood-brain barrier (BBB,

dysfunctional in AD), elevating brain levels of chemokines, Complement System (CS) factors and cytokines (increased in AD) (Tremblay et. al, 2020) that damage neurons (Gordon et. al 2021). TMPRSS2, ACE2 and AT1R Hippocampal expression levels correlate positively with severe AD (**Figure 2.3D**).

NFKB1 and RELA belong to the same Severe AD greenyellow Hippocampal module with many immune enrichments like: microglia, PID IL-1 Pathway, abnormal innate immunity and immunoglobulin level, innate immune system response, and activated NFKB signals survival (**Figure A.2** (SaniyaKhullar 2024)). In our GRNs, NFKB1 and RELA regulate genes of several cytokines associated with the severe Covid-19 Cytokine Storm and/or reduced long-term potentiation (**Figure A.7A-D** (SaniyaKhullar 2024)). We used our GRNs to analyze how NFKB1 and RELA regulate genes of several pro-inflammatory cytokines that are involved in the severe Covid cytokine storm [associated with BBB dysfunction, anti-neuron antibodies, neuroinflammation, neurodegeneration (Erausquin et. al 2021), activation of microglia and astrocytes]. In the DLPFC, RELA binds to an enhancer of *CXCL10*, which has altered levels associated with immune dysfunction and inflammatory disease severity (Liu et al. 2011). NFKB1 binds to LTL enhancer of *CSF3R*, a regulator of neutrophil (innate immune system cells that change level and function in severe Covid (Reusch et. al, 2021)) and microglia maintenance (Hampel et. al 2020). NFKB1 regulates IL-2 binding to an IL-2 enhancer on chromosome 4: 121,696,658–121,696,872. In AD brains, Aβ stimulation may activate NF-κB TFs to upregulate *TNFa* and *IL-1B* (regulates amyloid precursor protein (APP) synthesis) in microglia and astrocytes (Jha et. al 2019), likely triggering neuron death, cytokine cascade, more plaques, inflammation, tissue destruction (Kinney et al. 2018b; Landhuis 2021). NFKB1 and RELA regulate TFs that further regulate inflammatory cytokines *IL-1B, IL-12B, CCL2, MMP1/3, CLGN*. In the Hippocampus, NFKB1 regulates SPI1 and BATF that then jointly regulate *MMP1*. RELA regulates TNFa-induced proteins TNFAIP3/6 (regulate long-term potentiation in AD) in the Hippocampus; IL-2 and TNFa are highly expressed in Covid patients with severe pneumonia who develop ARDS (needing to go to the ICU and receive emergency oxygen (Kircheis et al. 2020)) as well as in severe AD patients (as are cytokines *CCL2, IL-1B, IL-12B*). RELA activates IL-12A/B (recruits Natural Killer cells (Roberts 2015)) and IL-1B via their respective enhancers and regulates IL-6 (induces

C-Reactive Protein (CRP) synthesis that activates the Complement System (Zass et al. 2017)) by binding to an IL-6 promoter.

Activation of the immune-related Complement System (CS) is involved in an inflammatory feedback loop with neutrophil activation, resulting in tissue injury (Java et. al, 2020) in severe Covid. CS components like the C1qrs enzyme complex activate microglia to the M1 state, releasing inflammatory mediators (Java et. al, 2020) causing Hippocampus atrophy (Marshe et. al, 2021) (these M1 microglia induce neurotoxicity; M2 microglia are instead anti-inflammatory and neuroprotective). We found that C1qrs correlates negatively with Control and Initial AD, but positively with AD severity (e.g. Moderate and Severe AD) in the Hippocampus (**Figure A.8** (SaniyaKhullar 2024)). Indeed, many CS components correlate positively with AD progression (**Figure A.9** (SaniyaKhullar 2024)). APOE genotype is also associated with differences in Complement Cascade Component C1qrs expression in Covid-19 patients (Inal 2020) in the DLPFC (negative correlation with APOE E2 allele, positive with the APOE E4 allele) (**Figure A.10** (SaniyaKhullar 2024)). Immunoglobulin-G (IgG) antibodies, whose responses to epitopes are key to Covid (Heffron et. al, 2021; Ong et al. 2024) immune response, correlate positively with moderate but not severe AD. Fibrinogen and the SELP protein changed from negative to positive associations from moderate to severe AD. **Figure A.11** (SaniyaKhullar 2024) shows correlations between other shared AD–Covid mechanisms (based on KEGG pathways) and AD phenotypes (e.g. Hippocampus: Tumor Necrosis Factor Receptor (TNFR) with severe AD, IkappaB kinase (IKK) with cognitive impairment; LTL: IKK with neuritic plaques; DLPFC: c-Jun N-terminal kinases (JNKs) with cognitive resilience). We identified other KEGG AD-COVID pathways that are highly correlated with having AD (**Figure A.12A-C** (SaniyaKhullar 2024)) across the 3 brain regions.

By targeting GRNs related to the shared AD-Covid pathways like the NFKB pathway, we can uncover new therapeutic targets and strategies to manage both AD and COVID-19. Modulating the NFKB pathway could help reduce neuroinflammation and slow disease progression of AD, and can mitigate the hyperinflammatory response, potentially reducing the severity and improving outcomes for patients with severe Covid-19. For Covid-19-induced cognitive impairments (that are similar to AD in

terms of symptoms), understanding the NFKB pathway's role can help address the long-term cognitive impacts seen in Covid-19 survivors. Exploring the commonalities in inflammatory pathways, particularly the NFKB pathway, may help uncover new therapeutic targets and strategies to address both Covid-19-induced cognitive impairments and AD.

Figure 2.3 – Gene regulatory networks and phenotypes for NFKB, a shared pathway of AD & Covid-19.

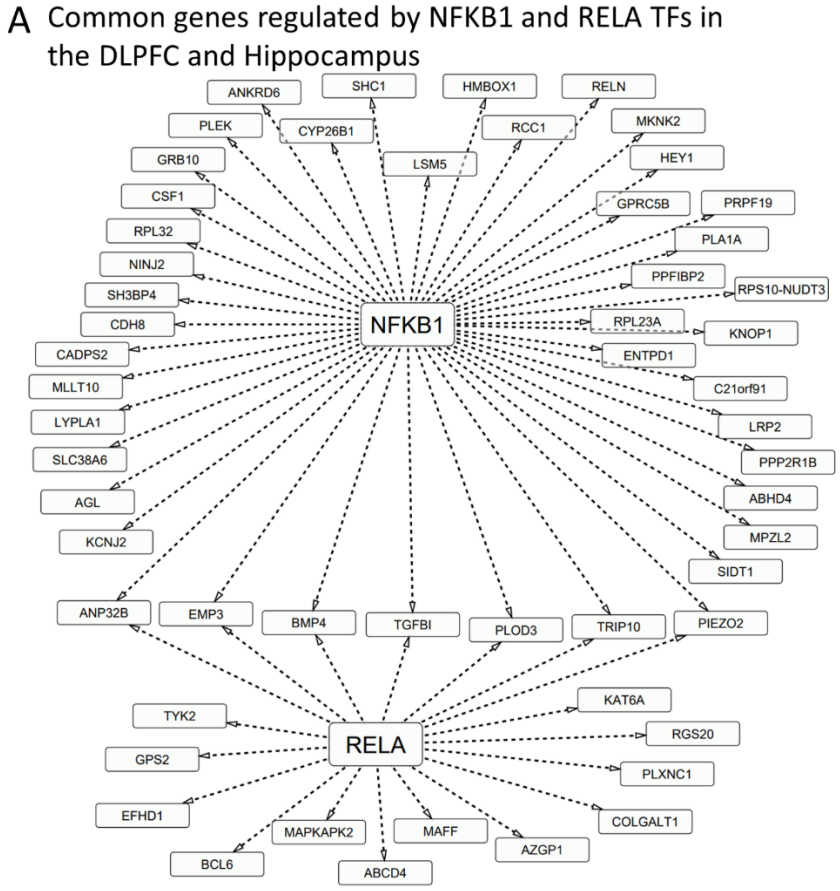


Figure 2.3A) – A subnetwork focusing on overlaps in the GRN between the DLPFC and Hippocampus, focusing on the target genes (TGs) regulated by NF-κB TFs: RELA (belongs to NF-κB class II) and NFKB1 (belongs to NF-κB class I). Here, only TF–TG links found in both brain regions are shown.

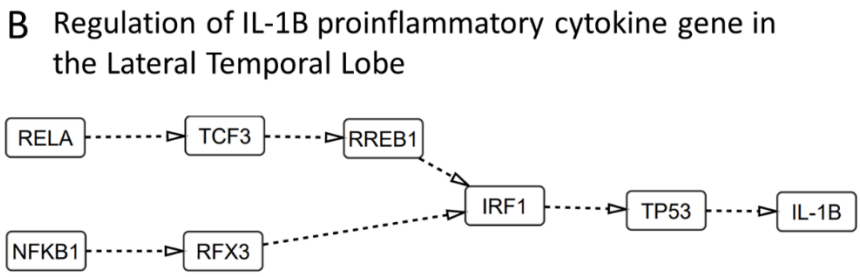


Figure 2.3B) – RELA and NFKB1 TFs regulate other TFs in a domino chain reaction that then regulate the pro-inflammatory cytokine *IL-1B* in the LTL.

This illustrates the complexity of GRNs. For instance, RELA regulates TCF3, which then regulates RREB1, which then regulates IRF1, which then regulates TP53, which lastly regulates *IL-1B*.

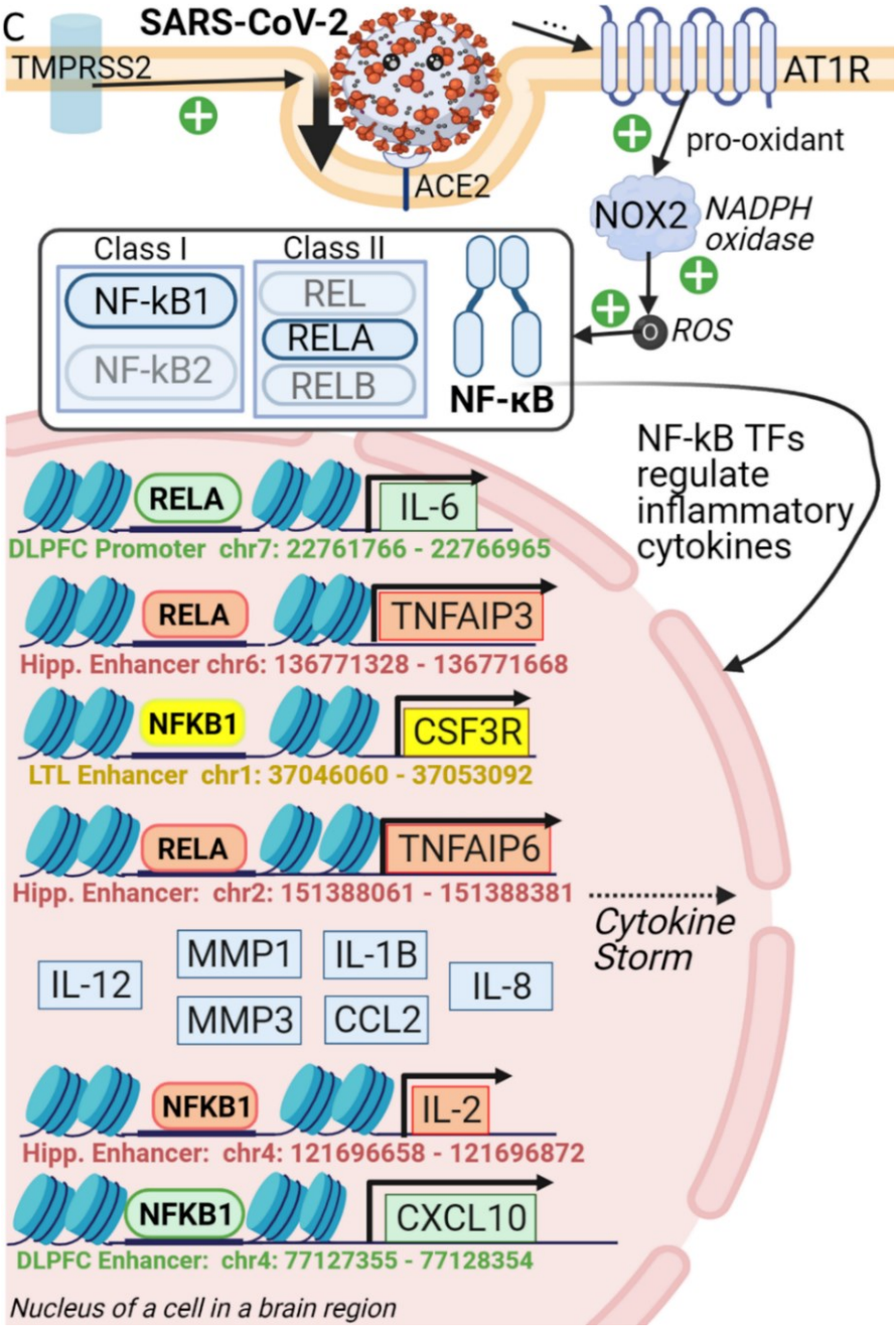


Figure 2.3C) – Gene regulatory networks & phenotypes for NFKB, a shared pathway of AD & Covid-19.

The Covid-19 virus (SARS-CoV-2) spike protein enters and infects the cell. Gene regulation of pro-inflammatory cytokines by activated NF-κB TFs from our Hippocampal, LTL and/or DLPFC GRNs is linked with severe Covid-19 outcomes (e.g. cytokine storm and beyond). This visualization is adapted from the

Covid-19 KEGG network (hsa05171), focusing on the NF-κB pathway. Gray dashed arrows indicate regulation and black arrows indicate activation of cytokines by the respective TF. GRN edge lists in **§A.4** (**File A4** for the Hippocampus, **File A5** for the LTL) (SaniyaKhullar 2024) show more examples.

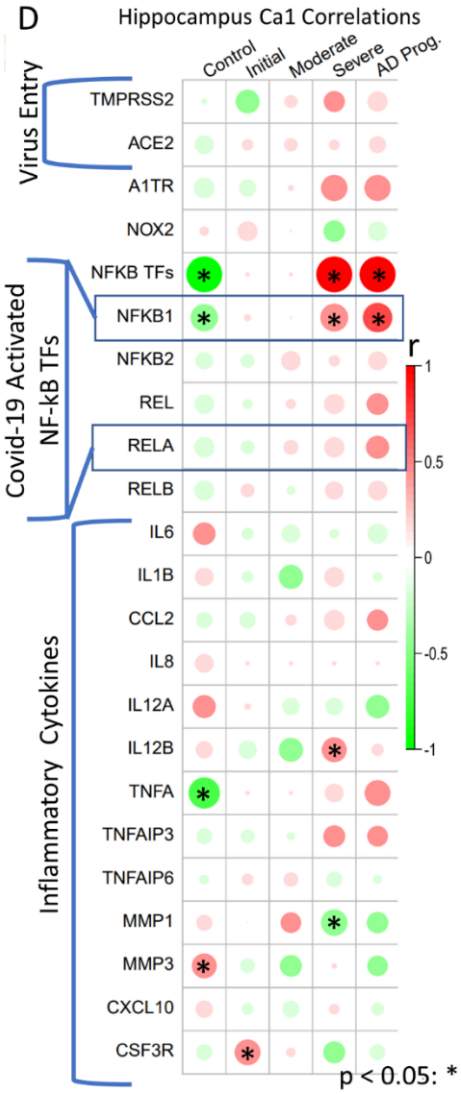


Figure 2.3D) – Pearson correlations between AD phenotypes and expression levels of genes from **Figure 2.3C** in the Hippocampus; the gene-phenotype correlations with *P*-value < 0.05 are denoted with an asterisk (*) on top.

2.3.5 Machine learning prediction of Covid-19 severity from AD-Covid related gene regulatory networks (GRNs)

There are several other shared AD–Covid mechanisms. **File A6** (SaniyaKhullar 2024) and **File A7** (SaniyaKhullar 2024) have results for this section. Covid-19 phenotype positively correlates with many AD KEGG mechanisms (**Figure A.13** (SaniyaKhullar 2024)). We normalized gene expression data of a

recent Covid-19 cohort ($N = 50$ Intensive Care Unit (ICU) vs. $N = 50$ non-ICU human samples) (Overmyer et. al, 2021) (**Figure A.14** (SaniyaKhullar 2024)) and identified 5,085 differentially expressed genes (DEGs, 2,505 upregulated and 2,580 downregulated) for severe Covid (ICU). We looked at our three final brain region GRNs related to the 22 shared genes between AD and Covid KEGG networks, including TFs that regulate them and/or their TGs i.e. AD–Covid GRNs (**Table A.6** (SaniyaKhullar 2024)). We then identified the AD–Covid lists from these 3 AD–Covid GRNs and filtered these gene lists down to only include the genes from the modules associated with AD phenotypes (21, 28, and 35 modules for the Hippocampus, LTL and DLPFC, respectively). Finally, a combined AD–Covid gene list had 2,153 genes (pooling our final AD–Covid genes lists from the Hippocampus: 1,146 genes, DLPFC: 895, LTL: 322) (**Figure A.15** (SaniyaKhullar 2024) details this process) of which 733 are severe Covid DEGs. Covid-19 severity correlates positively with many AD KEGG mechanisms and vice-versa (**Figure A.12**-**A.13** (SaniyaKhullar 2024)). Seven DEGs are in all 4 gene lists (5 upregulated genes like *SPI1*, 2 downregulated genes: *PIK3R3* and *STAT2*). AD–Covid genes strongly associate with Covid-19 severity.

We applied support vector machine (SVM or SVC, **§2.2: Materials and Methods**) models to predict the probability of severe Covid-19 outcomes in Covid patients. Each model was trained using this normalized Covid gene expression data for a list of genes. We applied recursive feature elimination (RFE) cross validation (CV) on an SVM model for the 18 benchmark Covid genes (from previous studies (Hu et. al, 2021; Pairo-Castineira et. al 2020; Hou et al. 2020; Kong et al. 2020)) (**Figure 2.4A**); RFECV found 10 benchmark genes were optimal (highest 5-fold stratified CV accuracy on training data). We ran RFE on each of our 5 lists (benchmark and 4 AD–Covid lists) to select the top 10 optimal genes (predictive for Covid severity) for each list (based on the training data), which we then used to build our benchmark model and 4 AD–Covid models, respectively. Forty-six genes were found across all 5 input lists (benchmark, combined, Hippocampus, LTL, DLPFC); the 36 genes from our four AD–Covid models are our AD–Covid genes (we found 0 overlaps with the 10 benchmark genes). Our 4 AD–Covid models outperformed the benchmark model on training data with higher average area under the receiver-operator characteristic curve (AUC) (**Figure A.16** (SaniyaKhullar 2024)) and accuracy (**Table A.7** (SaniyaKhullar

2024)). Our models perform better than the benchmark model on test data (20 balanced samples) with higher AUC (except for the LTL model) and accuracy (**Figure 2.4B**). Relative to the benchmark model, the DLPFC model (optimal; accuracy: 85%, AUC: 0.98) boosted accuracy by 25% and AUC by 0.19. Decision curve analysis (DCA(Vickers and Elkin 2006; Sørensen et. al 2018), **§A.1** (SaniyaKhullar 2024)) found that our models generally have higher clinical Net Benefits than the benchmark model across all probability thresholds (from 0% to 100%; average Net Benefit increase of 0.153) and therefore have a greater clinical usability (**Figure A.17** (SaniyaKhullar 2024)). Hence, using our optimal AD–Covid model (for a given probability threshold) on average increases the number of truly severe Covid patients detected by approximately 153 per 1,000 Covid patients, without changing the number of non-severe patients who are needlessly sent to the Intensive Care Unit (ICU). Overall, our 36 genes have higher predictability for Covid severity than benchmark Covid genes on new Covid patient blood gene expression data. Our AD–Covid models may provide potential novel strategies to guide clinical decisions on sending Covid patients to the ICU or not.

We found that our 36 AD–Covid genes (**Table A.8** (SaniyaKhullar 2024)) driving Covid severity may also drive neuroinflammation, which is predictive of AD. For this, we trained a logistic regression (LR) model to predict the probability of AD using Superior Frontal Gyrus (SFG) brain region gene expression data. Three AD–Covid genes (*ANP32B, GPI, SPI1*) are AMP-AD (Ryan and Petanceska et. al, 2022) nominated AD genes. *GPI* promotes neuron survival and immune-functions [e.g. serves as tumor-secreted cytokine (AMP-AD (Agora))]. TF SPI1 regulates immune functions and microglia-mediated neurodegeneration in AD (Rustenhoven et al. 2018a), and correlates strongly with AD/Braak Progression in the Hippocampus. We used 35 of those 36 genes (*SMIM27* was missing) and added four binary (dummy: 0/1) features to control for the four cell-types (39 features in total). We compared our test performance for 24 samples: (12 AD, 12 Control) with that of a LR model using 597 AMP-AD (Ryan and Petanceska et. al, 2022) genes (601 features). Our AD–Covid LR model outperformed the AMP-AD LR model: AUC (0.583 vs. 0.569), accuracy (70.3% vs. 62.5%), DCA (29 optimal probability thresholds vs. 21) (**Figure A.18A-B** (SaniyaKhullar 2024)).

Thus, our 36 AD–Covid genes are predictive of not only Covid severity but also of AD, as they performed better than their respective benchmark models and have promising clinical translational ability for predicting immune dysregulation, inflammation, AD and severe/neuro Covid. Gene *ANP32B* [enriched in extracellular vesicles in AD mice brain tissues (Muraoka et. al, 2021)] strongly predicts Covid severity as it was found in DLPFC and Hippocampus models; *ATM, EMP3*, and *LILRA6* were found in Hippocampus and combined models. **Figure A.19** (SaniyaKhullar 2024) reveals how 13 of these 36 genes are DEGs in excitatory (ExNs) and/or inhibitory (InNs) neurons for AD pathology overall and/or early AD pathology versus none using recent data (Mathys et. al 2019); for instance, *SPI1* is downregulated in ExNs in both comparisons, whereas *MYLIP* is upregulated in InNs for early AD (versus controls). We highlight the DLPFC GRN subnetwork for all 10 predictive genes directly regulating or regulated by at least 1 of the 22 shared KEGG genes (**Figure 2.4C**, Hippocampus/LTL: **Figure A.20A-B** (SaniyaKhullar 2024)). Our 3 brain region GRN subnetworks reveal TF–TG interactions that may predict pro-inflammatory cytokine levels and neuroinflammation. NFKB1 and RELA regulate several genes across all three regions, associated with immune dysregulation (able to predict Covid severity), like *ANP32B* in the DLPFC. SPI1 and NFKB1 target *PLEK*, whose expression is linked to synapse failure and cognitive dysfunction in AD (Guo et al. 2019). STAT5B regulates glucocorticoid receptor activity, which impacts the expression of pro- and anti-inflammatory genes (Zass et al. 2017). We found that STAT5B jointly regulates PI3K subunits *PIK3CD* and *PIK3CB* in the DLPFC (NFKB2 regulates *PIK3R1* in the LTL); altered PI3K (shared AD–Covid mechanism) signaling may increase *IRF5* activity (Naughton et al. 2020) in AD (Gabbouj et. al 2019) and in severe Covid. These 10 DLPFC SVC model-based AD–Covid genes are enriched (Zhou et al. 2019) with immune system diseases like Hodgkin Lymphoma, T-cell Leukemia, Waldenstrom Macroglobulinemia.

Figure 2.4 – Prediction of Covid-19 severity using AD-Covid gene regulatory networks (GRNs).

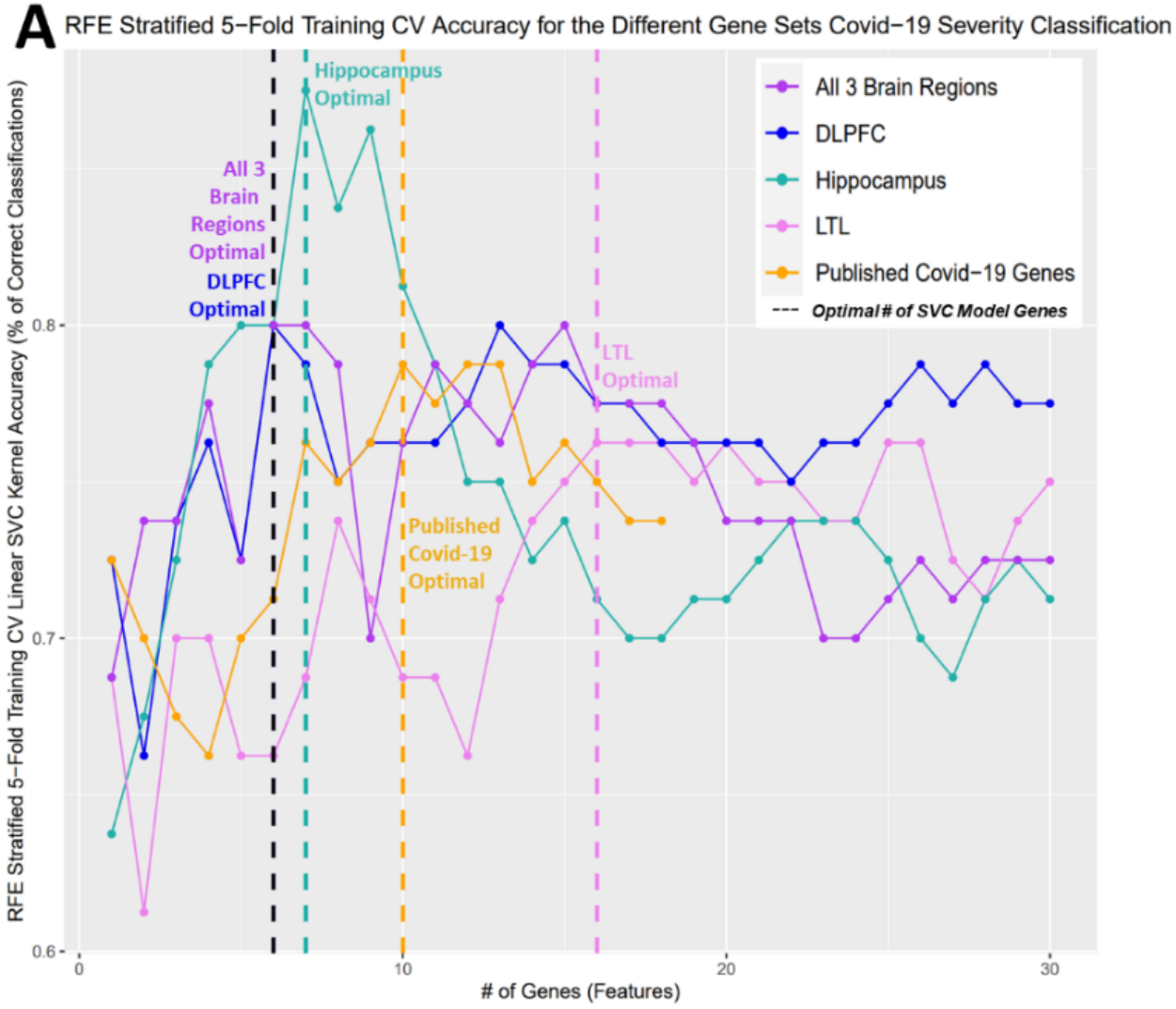


Figure 2.4A) – Prediction accuracy of Covid-19 severity after selecting different numbers of genes from AD–Covid GRNs and recently found Covid-19 genes (benchmark genes). The accuracy was calculated based on the support vector machine classification (SVM or SVC) model with 5-fold stratified cross-validation on 80 balanced training samples. The dashed lines correspond to the minimal numbers of select genes with the highest accuracy (i.e. optimal gene sets for predicting Covid-19 severity).

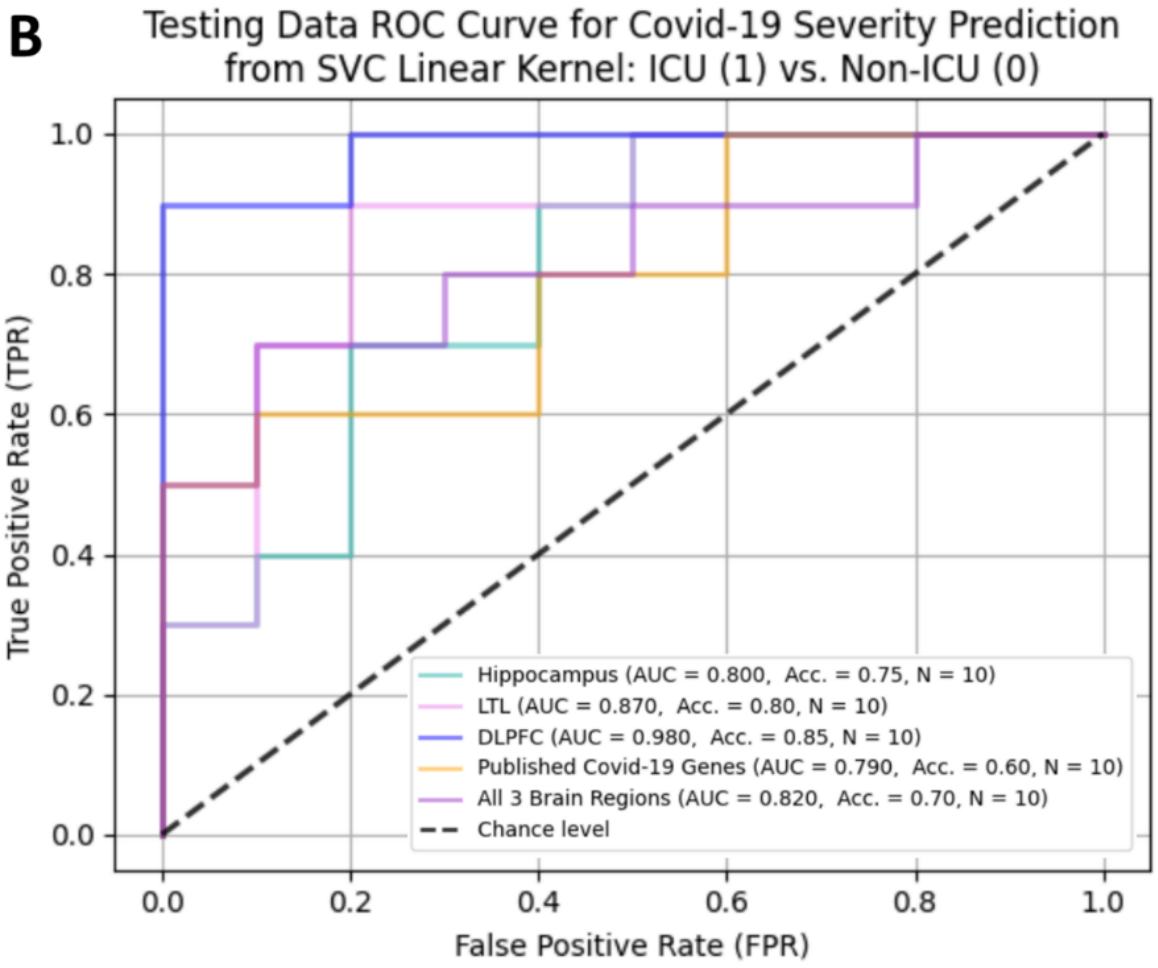


Figure 2.4B) – Receiver operating characteristic curves and corresponding AUC values for classifying Covid-19 severity in the test data of 20 balanced samples using the SVC machine learning models.

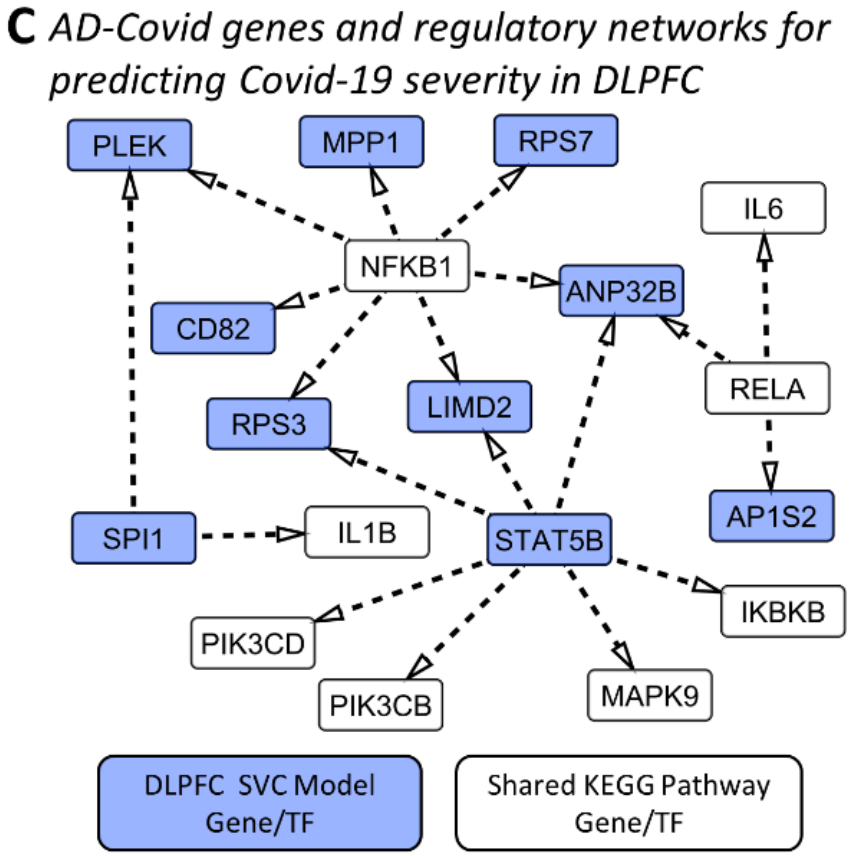


Figure 2.4C) – Subnetwork of the DLPFC GRN relating to the 10 AD–Covid DLPFC genes for predicting Covid-19 severity ($N = 10$) with the shared KEGG genes. Blue: genes/TFs found in the optimal DLPFC final model (which are also 10 of the 36 AD–Covid genes). White: 1 of the 22 shared KEGG genes (between AD and Covid KEGG networks). There is no overlap between both sets of genes.

2.3.6 Identification of disease risk variants for AD phenotypes via integration of GWAS and gene regulatory networks

It is crucial to understand how non-coding disease-associated SNPs (over 90% of risk SNPs (Kumar et al. 2017)) affect gene regulatory mechanisms that eventually impact AD phenotypes. For this purpose, we looked at both AD SNPs and Covid-19 severity (based on hospitalization status upon Covid infection) SNPs from recent GWAS. We did this for two main reasons. First, studies have found that even mild Covid-19 infection is associated with brain changes (Abbasi 2022) and that severe Covid SNPs may contribute to cognitive dysfunction (Gordon et. al 2021), thereby worsening AD phenotypes. Second, incorporating severe Covid SNPs may help us discover how Covid-related genetic risk variants are

associated not only with Covid severity but also with AD and cognitive impairment (e.g. neuro-Covid), both of which are currently unknown (Gordon et. al 2021). We mapped AD SNPs and Covid severity SNPs onto our final GRNs to see how these SNPs alter TFBSs on regulatory elements (enhancers and/or promoters) that regulate target genes (TGs) and gene modules. Furthermore, we linked these SNPs to AD phenotypes of corresponding TGs and modules for our three brain regions i.e. 'brain-region SNP-effected-GRN for AD phenotypes': SNP–TF–Regulatory Element–TG–Module–Phenotype. Thus, we could predict how AD and/or severe Covid SNPs impact TF regulation of TGs that belong to modules enriched with biological functions; TGs and modules may associate with AD phenotypes. Our SNP-effected-GRN predicted 144,098 total unique SNP–TF–TG relationships across the three regions (for 17,795 SNPs impacting TFBSs, 14 common AD–Covid SNPs); 6,245 SNP–TG relations had at least 1 validated blood/brain expression quantitative trait loci (eQTL) link. **File A8** (SaniyaKhullar 2024) has metrics and our annotated SNP-effected-GRN. Below, we highlight strong examples from our many SNP-effected-GRN predictions.

Our SNP-effected-GRN may predict how AD and/or severe Covid SNPs may alter the expression of TGs like our 36 AD–Covid genes. In **Figure 2.5** we use our Hippocampus SNP-effected-GRN to focus on *NFYA* (Nuclear Transcription Factor Y Subunit Alpha), 1 of 525 common TGs dysregulated by AD SNPs and by severe Covid SNPs. **Figure 2.5A** shows many predicted *NFYA* enhancers. AD SNP rs2073014 strongly hinders EHF and ELF1 TFs (that both belong to the ETS TF family (Corces et al. 2020)) from binding to an *NFYA* enhancer. On the other hand, severe Covid SNP rs2495242 strongly increases HSF2 regulation of *NFYA*. SNP rs2073014 is a mutation that changes the DNA base from a T to a C at chromosome 6 position: 41,029,109, disrupting EHF and ELF1 motifs (**Figure 2.5B**); both motifs are significantly enriched in all single-cell assay for transposase-accessible chromatin (scATAC-seq) peaks (typically corresponding to enhancers) (Corces et al. 2020) of open chromatin in microglia. We found 48 prefrontal cortex (PFC) eQTL SNPs associated with reduced NFYA expression that correlate positively with rs2073014 based on linkage disequilibrium (LD) analysis; this verifies that rs2073014 is associated with *NFYA* expression in the brain. Rs193235873, a Covid severity SNP, increases TP63

regulation of *E2F4* (and TP63 significantly regulates *E2F4*'s Braak 6 stage module), which in turn regulates NFYA (**Figure 2.5C**). NFYA is associated with AD (Gupta et al. 2022), plays a key role in various cancers (Li et al. 2020) and regulates 5 AD–Covid genes like *LILRA2* and *ANP32B* (their shared module positively correlates with worse AD phenotypes). **Figure A.21A** (SaniyaKhullar 2024) identifies cell-type eQTL SNPs impacting the regulation of *ANXA11* (an AD-Covid gene) by five TFs, like NFKB1, in the Hippocampus. In particular, SNPs linked to our AD–Covid genes and correlated with AD phenotypes may help explain genetic mechanisms of critical illness, neuroimmunology and cognitive impairment in Covid (Pairo-Castineira et. al 2020) and in AD.

We predicted 5 shared AD and severe Covid SNPs in *IFNAR2* Hippocampus and/or LTL enhancers that may impact regulation of *IFNAR2*, a known Covid severity gene (Jalkanen et al. 2023). These 5 SNPs are in LD with blood eQTL SNP rs7509997 that is strongly positively associated (P = 3.31e-49) with *IFNAR2* expression. During AD, cytokine *CSF3R*'s overexpression in the LTL may be partly explained by SNPs like rs483341 that disrupt the ability of TF REST (protects neurons from Aβ-protein toxicity (Lu et. al 2014)) to bind to chromatin to repress its TGs (Maezawa et. al 2012) like *CSF3R*, leading to inflammation (Magno et al. 2019b). Harmful AD SNP rs2564970 (P = 5.47e-08), which is 4 bases from a predicted *CR1* (Complement receptor type 1) Hippocampus enhancer (chromosome 1: 207,464,045 - 207,464,283), may strongly disrupt NFKB1 and RELA regulation of *CR1*, a major AD gene associated with the complement system (CS). Moreover, our SNP-effected-GRN predicts previously unknown SNPs and specific TFs associated with NF-κB TF activation in AD, which may make NF-κB TFs neuroprotective or neurotoxic. We predict that harmful Covid severity SNP rs2736322 disrupts RREB1's ability to bind to an *FAM167A* LTL enhancer and subsequently regulate *FAM167A*, a TG which correlates positively with AD and Braak stages and belongs to an AD LTL module. This may explain this SNP's negative eQTL relationship with *FAM167A* expression across various brain cell-types. Furthermore, we have **Figures A.21B-C** and **A.22A-B** (SaniyaKhullar 2024) that elaborate on the following select stories: AD SNPs on regulatory elements may dysregulate Hippocampus expression of 3 AD–Covid genes: *EMP3*, *LILRA2*, *SPI1* (**Figure A.21B** (SaniyaKhullar 2024)). Harmful AD SNP rs754366 (has

a positive Prefrontal Cortex (PFC) eQTL link to APOC2 expression) may increase SPI1 binding to an *APOC2* DLPFC enhancer where it activates *APOC2* (**Figure A.21C** (SaniyaKhullar 2024)).

Non-coding AD SNPs in microglia scATAC-seq peaks may impact regulation of *KCNN4* (Potassium Calcium-Activated Channel Subfamily N Member 4), an AD risk microglia gene with previously no known mutations (Maezawa et. al 2012) associated with alterations in its expression (**Figure A.22A** (SaniyaKhullar 2024)). Different Hippocampal and LTL SNPs impact regulation of *KCNN4*, a key AD drug target overexpressed during AD. We visualize the impact of rs62117780 on FOXC2 and POU2F2 regulation (and *KCNN4*'s Braak progression module) in the Hippocampus and rs4802200 on E2F7 regulation in the LTL. *KCNN4* belongs to the magenta AD LTL module that has key enrichments like death receptor signaling, autoinflammatory disorder, TNFa/NFkB Signal Complex (**Figure A.2B** (SaniyaKhullar 2024)). Hence, increased *KCNN4* expression is associated with AD progression in both regions. Rs62117780 is in a microglia signal peak (**Figure A.22B** (SaniyaKhullar 2024)), consistent with findings that *KCNN4* is mainly expressed in microglia and regulates microglial activation by modulating Calcium (Ca2+) influx signaling and membrane potential (Maezawa et al. 2012). *KCNN4* has low expression in healthy neurons and is associated with neuroinflammation and reactive gliosis during AD. Blocking *KCNN4* likely curbs microglial neurotoxicity, leading to slower neuronal loss and better memory levels (Yi et al. 2017). This link uncovers how AD SNPs regulate *KCNN4* expression in AD.

Next, we focused primarily on AD-related GWAS SNPs. We found many regulatory networks associating various non-coding AD SNPs with AD phenotypes (interpretation guide: **Figure A.23** (SaniyaKhullar 2024)). Low *SPI1* expression in Hippocampus controls may reduce microglial-mediated neuroinflammatory responses and delay AD onset (Rustenhoven et al. 2018a). SPI1, a microglia master regulator TF, regulates immune functions in AD (Yashiro et al. 2019), is strongly correlated with AD (correlation $r = 0.355$), AD Progression ($r = 0.375$), Braak progression ($r = 0.437$), and Braak 6 stage ($r = 0.407$), and belongs to a severe AD gene co-expression module ($r = 0.41$). In the Hippocampus, many SNPs may disrupt the ability of various TFs to regulate *SPI1* (**Figure A.24A** (SaniyaKhullar 2024)); as a TF, SPI1 significantly regulates *DMPK* and its Severe AD Hippocampal module, and its regulated genes

are upregulated in microglia, leading to microglia-mediated AD neurodegeneration (Rustenhoven et al. 2018b). SNP rs2834164, which is associated with AD and with Covid-19 severity, may disrupt the ability of POU4F2 to regulate *IL10RB* in the Hippocampus; we find eQTL support for this SNP relationship and *IL10RB* is associated with AD and belongs to a Control Stage gene module that is positively correlated with better performance on the Mini-Mental State Exam (**Figure A.24B** (SaniyaKhullar 2024)). In **Figure A.24C** (SaniyaKhullar 2024), we visualize some AD risk SNPs that alter the regulation of target gene *ACE* in the Hippocampus and the LTL brain regions.

Further, we visualize how Covid-19 severity risk SNPs alter the regulation of *CCR1*, an early and specific marker of AD (Halks-Miller et al. 2003) across all 3 brain regions (**Figure A.24D** (SaniyaKhullar 2024)). In the Hippocampus, *CCR1* belongs to the Braak 2 associated gene module and is associated with the Braak 4 stage (mild dementia), while it is associated with Braak 1 or 2 stages (asymptomatic outcomes) in the LTL. The DLPFC sees increased expression of *CCR1* associated with AD, Braak progression, cognitive impairment, frequent neuritic plaques, dementia overall, along with other AD progression phenotypes. Similarly, *CCR1* belongs to a severe stage module in the DLPFC. Chemokines like *CCR1* are secreted by astrocytes and play core roles in AD pathology and neuroinflammation (Liu et. al 2014). Additionally, a previous study (Halks-Miller et al. 2003) found that CCR1-positive plaque-like structures in the hippocampus and entorhinal cortex are strongly associated with dementia severity and specifically correlate with amyloid beta peptides of the 1-42 species (Abeta42)-positive neuritic plaques in AD; importantly, examination of seven other dementing neurodegenerative diseases revealed that CCR1 immunopositivity was absent unless Abeta42-positive plaques were present. These findings highlight that neuronal CCR1 does not act as a general marker of neurodegeneration but is instead a component of the neuroimmune response to Abeta42-positive neuritic plaques. Further, recent studies have identified severe Covid-19 risk SNPs colocalize with regulatory elements and alter chemokine receptor gene control in monocytes and macrophages; in fact, such SNPs have been linked to increased expression of chemokines like *CCR1* in inflammatory monocytes and macrophages and higher risk of hospitalization post-Covid infection (Stikker et al. 2022). Our SNP-effected-GRNs thus underscore the role

of an immune system gone haywire in not only severe Covid-19 outcomes but also in AD progression phenotypes across multiple brain regions.

Figure A.25A-B (SaniyaKhullar 2024) shows how 1 AD risk SNP, rs3851178, may alter TF binding abilities and subsequent regulation of different TGs across brain regions. This example shows how the SNP-effected-GRNs enable comparative analysis of regulatory networks and phenotype outcomes across brain regions.

In **Figure A.26A-B** (SaniyaKhullar 2024), we examine the varying impact of AD risk SNP rs78073763 on gene regulation of *PPP1R37*. In the LTL, *PPP1R37* is positively associated with the Control stage; this gene is associated with a severe cognitive impairment gene module in the DLPFC and is positively correlated with Braak progression and with various AD-related phenotypes. This SNP changes the DNA from a T to a G (at chromosome 19: 45,649,838 position, hg19 genome) and may boost RARA regulation of *PPP1R37* in the LTL, disrupt PAX5 and SPIC regulation of *PPP1R37* in the DLPFC, and boost GCM1's regulation of this gene in the DLPFC. *PPP1R37* expression is strongly associated with *APOE* expression with extensive cross-tissue effects on AD (Liu et al. 2021). As non-coding SNPs may have highly cell-type specific effects, we explored the epigenetic landscape and regulatory element signals for significant putative functional SNP rs78073763, that impacts expression of *PPP1R37* (co-expressed TG with *APOE*). This SNP is present in microglia-specific regulatory elements (Corces et al. 2020) (**Figure A.26C** (SaniyaKhullar 2024)), underscoring dysregulated microglia and neuroimmunology in AD.

Our pipeline flags candidate Covid-19 susceptibility non-coding SNPs, which may also worsen AD phenotypes, most likely via triggering a cascade of neuroinflammatory pathways. The above stories and more (**§A** (SaniyaKhullar 2024)) underscore the importance of our work and findings. Moreover, our SNP-effected-GRN may help explain GRN mechanisms behind several causal blood/brain eQTL links.

Figure 2.5 – Select SNP regulatory networks (SNP-effected-GRNs) linking AD and Covid-19 severity risk variants (GWAS SNPs) to AD phenotypes in the Hippocampus.

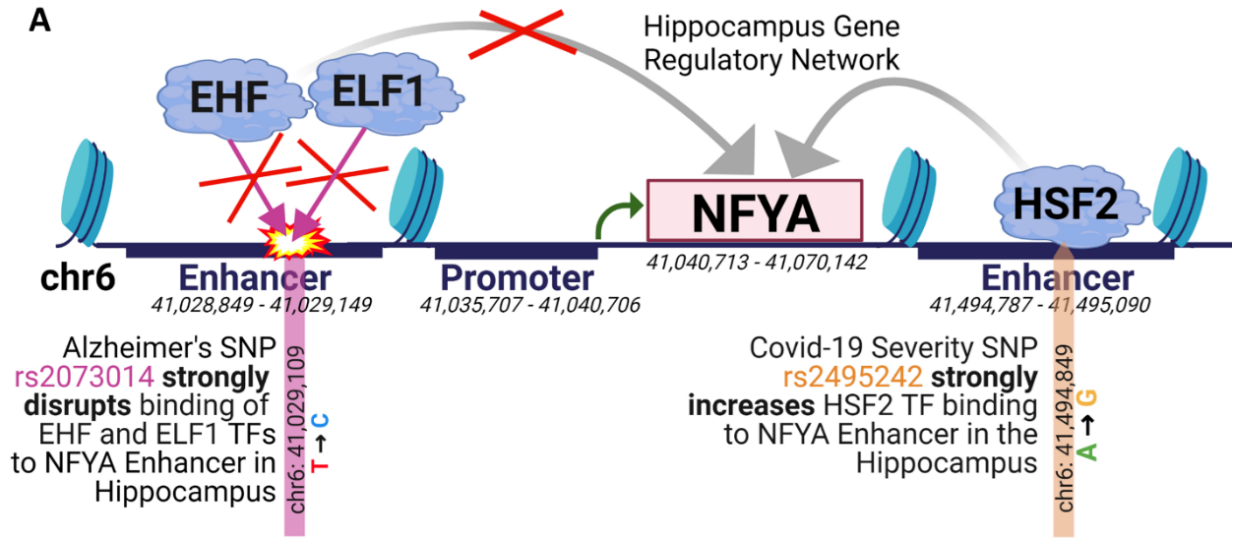


Figure 2.5A) - AD SNP rs2073014 may disrupt binding of 2 TFs (EHF and ELF1) to an *NFYA* enhancer. Another Covid-19 severity SNP instead increases HSF2 TF binding and subsequent regulation (activation or repression is unknown) of *NFYA*.

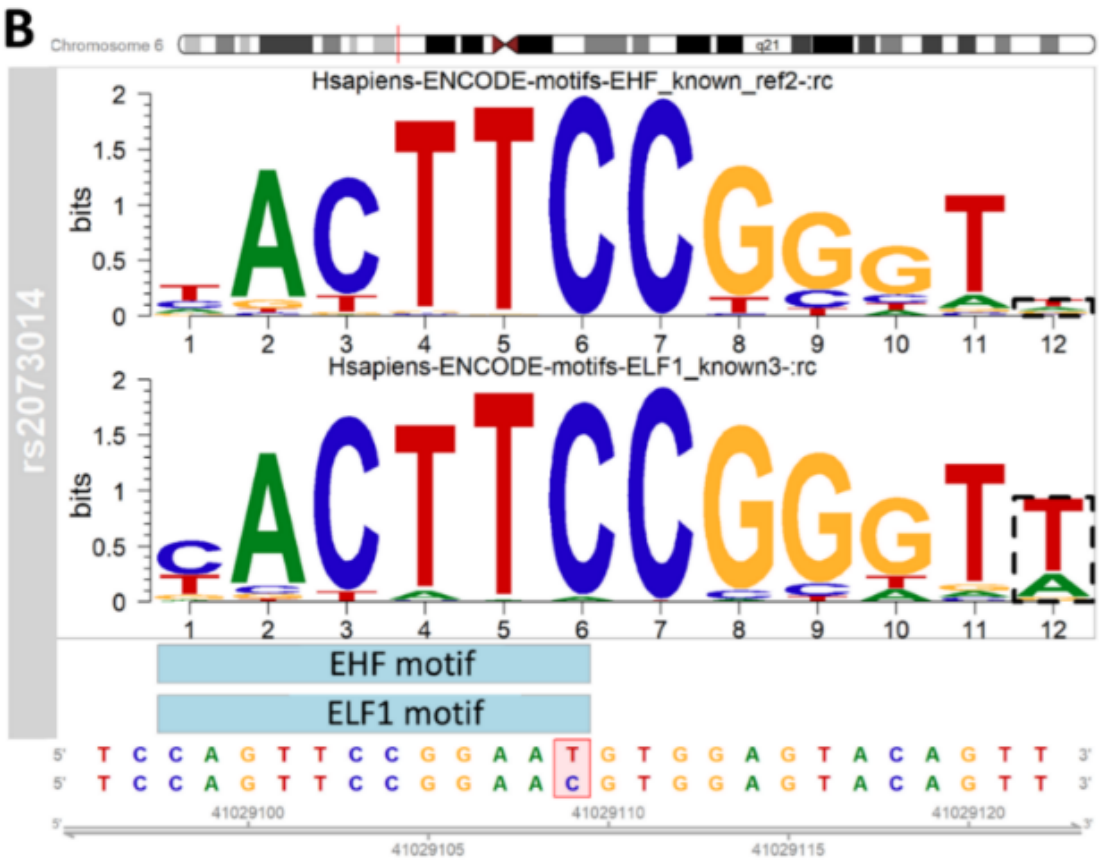


Figure 2.5B) - This shows how AD SNP rs2073014 interrupts the binding sites of EHF and ELF1 TFs in the Hippocampus (on the basis of their respective sequence-specific motifs).

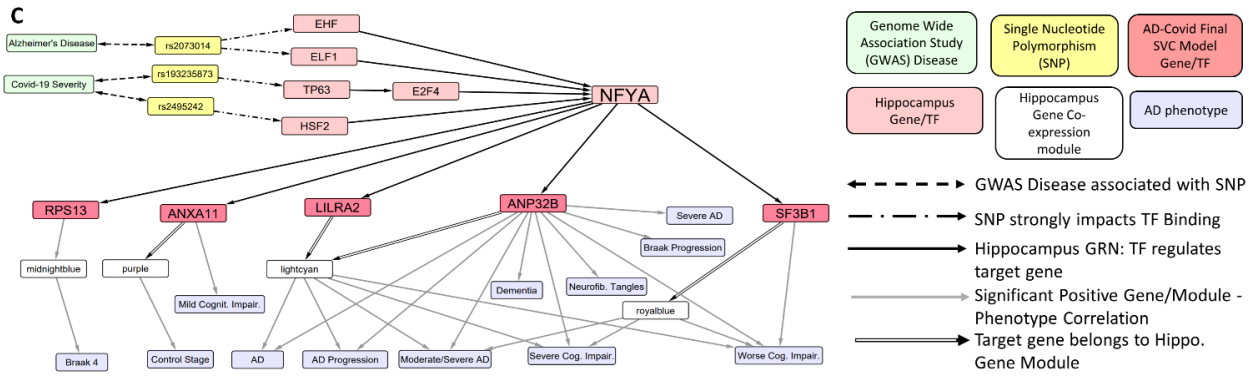


Figure 2.5C) - On the right is a legend for this network. Bi-directional dashed arrows represent SNP association with AD and/or severe Covid-19 on the basis of summary statistics from recent GWAS. Arrows with dots in the middle represent that the SNP strongly impacts TF Binding (either increasing or disrupting it; details available in **§A.4** (SaniyaKhullar 2024)). The solid arrow represents that the TF regulates that given TG in the Hippocampus GRN. Grey arrows represent that the TG and/or TG module are statistically significantly positively correlated ($P < 0.05$, $r > 0$) with that given AD-related phenotype. Here, we analyze the role of 1 AD SNP and 2 Covid severity SNPs in eventually impacting the regulation of *NFYA* in the Hippocampus. For instance, SNP rs193235873 impacts the ability of TP63 to regulate E2F4 that then regulates *NFYA*. Furthermore, we show how NFYA regulates 5 AD–Covid genes (*RPS13*, *ANXA11*, *LILRA2*, *ANP32B*, *SF3B1*) and we highlight the Hippocampus gene modules for these 5 genes (using black arrows that contain a white rectangle). In addition, we link the gene modules and these 5 AD–Covid genes with their respective AD-related phenotypes. Thus, we predict how AD SNPs and Covid-19 severity SNPs may eventually impact the regulation of 5 of our 36 AD–Covid genes that are associated with various AD phenotypes.

§ 2.4 Discussion

In the coming years, AD and Covid-19 will exert an ever-increasing toll on our society, making both diseases of paramount importance to address. Links among Covid, cognitive decline, and neurodegenerative diseases are puzzling and poorly understood (Gordon et. al 2021). Nonetheless, it is currently unknown whether Covid triggers new development of AD or accelerates progression of AD (Lindsey et. al 2022). There is on-going research on disabilities (e.g. memory, attention, sleep problems (Gordon et. al 2021)) in long-Covid (Lindsey et. al 2022). Individuals with neurological conditions like AD have high risk of Covid morbidity and mortality (Amruta et al. 2021). AD is linked to rogue immune mechanisms; pattern recognition receptors on astrocytes/microglia may respond to misfolded aggregated proteins, by releasing inflammatory molecules, worsening AD (Heneka et al. 2015).

In this chapter, we tried to investigate the role of neuroimmunology in AD and severe Covid-19, cognitive impairment associated with Covid, and AD-Covid links. Given the immense societal impacts of

AD and Covid-19, there is an urgent need for more research in this area. We performed an integrative multi-omics analysis (genotype, chromatin interaction, transcriptomics, epigenomics) to predict gene regulatory (GRNs) and gene co-expression networks in 3 AD-related brain regions. We used potential gene regulatory connections between AD and Covid-19 and AD-related gene co-expression modules for our machine learning analysis to predict Covid severity. We prioritized a set of 36 optimal 36 AD-Covid genes. Decision Curve Analysis (DCA) showed these 36 AD-Covid genes outperform known Covid-19 genes for predicting Covid severity (sending Covid patients to Intensive Care Unit (ICU) or not), demonstrating clinical translational ability of our predictive models. Further, those genes outperform nominated genes for predicting AD in a new brain region. Thus, our genes may be predictive of immune dysregulation and inflammation associated with Covid-19 and AD and can be targeted in follow-up studies. Lastly, we applied our SNPheno pipeline to build SNP Regulatory Networks (i.e. SNP-effected-GRNs) linking AD and/or severe Covid SNPs to our GRNs, co-expression modules, and AD phenotypes. We emphasized SNPs relating to our optimal 36 AD-Covid genes that may explain possible genetic causes for dysregulated neuroimmunology in AD and Covid. We flagged genes and SNPs for follow-up analysis. Our analysis can serve as a general-purpose tool to understand functional genomics and gene regulation in several other diseases.

Brain regions are composed of varied cell types that may impact co-expression networks and gene regulation; for example, AD patients may have fewer neurons and more immune cells. Many human brain cell-type GRNs were predicted from recent single-cell sequencing data (e.g. scRNA-seq, scATAC-seq), which enable studying cell-type functional genomics and GRNs (Jiang et. al 2020). In the future, our pipeline may be extended to single-cell transcriptomic data by applying a single-cell workflow for WGCNA: scWGCNA (Morabito et al. 2021), which constructs gene co-expression networks from snRNA-seq (single-nucleus RNA-seq: profiles mostly nuclear transcripts) and/or scRNA-seq (single-cell RNA-seq: profiles mostly nuclear and cytoplasmic transcripts (Bakken et. al 2018)) single-cell gene expression data (that suffers from inherent sparsity).

We validated that many phenotype-associated SNPs are located on regulatory elements with cell-type epigenomic activities. An integrative analysis of cell-type GRNs can also be performed to understand regulatory mechanisms for GWAS risk variants for refined AD phenotypes (e.g. cerebrospinal fluid, psychotic symptoms (Hampel et. al 2020)). Using pooled cell types in the Superior Frontal Gyrus (SFG) to predict AD may have confounding factors as a human sample could have many corresponding cell-type samples. That the Covid transcriptomic data was from blood samples may present limitations as AD–Covid GRNs use transcriptomic data from brain tissues. Still, researchers found immune dysregulation in both AD brain and blood samples (Guo et. al, 2019). Elevated pro-inflammatory molecules in Covid patients can compromise the blood-brain barrier (the BBB breaks down in AD), enter the brain and encounter astrocytes and microglia (both cell-types malfunction in AD); such patients are more susceptible to severe Covid and further neurological damage. Thus, Covid-19 patient blood gene expression data may predict future Central Nervous System (CNS) invasion and neuroinflammation. Our findings support further research into understanding better the causal links between AD and Covid (Kinney et al. 2018b) (e.g. it is unknown if Covid triggers new development of AD or accelerates AD progression). Treatments and drug development can perhaps be targeted at AD–Covid pathways to alleviate patient suffering i.e. care providers may suppress Interferon response or use acetylcholinesterase inhibitors (current AD treatment strategy) (Naughton et al. 2020) to stimulate the cholinergic anti-inflammatory pathway in Covid patients. Such treatments may reduce the overall risk of cognitive decline in Covid survivors (Gordon et. al 2021).

Currently, research in AD–Covid, long- or neuro-Covid and AD neuroimmunology is nascent and more data is being generated, which may be used to eventually expand on our current work (especially given sample size limitations of our data). Many large scientific consortia generate matched multi-omics data of individuals like AMP-AD (Ryan and Petanceska et. al, 2022), PsychENCODE (PsychENCODE Consortium), brainSCOPE (Emani et al. 2024), TCGA (Weinstein et al. 2013). We can extend our machine learning (ML) analysis to predict personalized phenotypes and prioritize phenotype-specific functional genomics and GRNs in diseases from this data. We found TFs regulate many gene modules and link to

AD phenotypes, suggesting possible collinearity driven by TF regulations across phenotypes. Emerging ML approaches like neural networks may decouple phenotypic collinearity, uncovering phenotypic-specific TFs. Studies emphasize systems biology and ML approaches (like ours) to identify biomarkers for neuroinflammation (Hampel et. al 2020) in AD, Covid-associated cognitive impairment (Gordon et. al 2021) (e.g. neuro-Covid or long-Covid) and Covid severity. Neuroimmunology research is discovering the role of dysregulated immune responses in other complex neurologic diseases like Schizophrenia (SCZ), Amyotrophic Lateral Sclerosis, Myasthenia Gravis, Parkinson's disease, Multiple Sclerosis (Coyle 2011). For instance, overactivated NF-κB signaling is also found in Post-Traumatic Stress (PTSD) and Bipolar (BD) Disorders (Zass et al. 2017). Covid-19 may similarly be used to understand the role of misguided immunity in SCZ, as SCZ is an autoimmune disease (excess pruning of synapses by microglia) and the second largest risk factor for Covid-19 death after age (NYU Langone Health, 2021; Nemani et al. 2021). Overall, we hope that our approach can be applied to help understand molecular mechanisms in other diseases by uncovering the association of orphaned GWAS loci in non-coding DNA regions with disease phenotypes and by using closely related diseases to help reveal additional mechanisms at play.

§ 2.5 Availability of data, software, and materials

Our analysis (codes and data including diagnosis information) is open-source available at **https://github.com/daifengwanglab/ADSNPheno** and our functional genomics resource for AD is available at **https://adsnpheno.shinyapps.io/AlzheimersDisease_SNPheno**. Our corresponding methods and materials, figures (**Figures A.1-A.26**), tables (**Tables A.1-A.8**), and data files (**Files A1-A8**) for SNPheno are available in **§ Chapter A** of the supplementary file (SaniyaKhullar 2024) that is hosted at: https://github.com/SaniyaKhullar/Supplementary_Chapters_Dissertation.

§ Chapter 3: NetREm: Network Regression Embeddings reveal cell-type transcription factor coordination for gene regulation

§ 3.0 Abstract

Transcription factor (TF) coordination plays a key role in target gene (TG) regulation via protein-protein interactions (PPIs) and DNA co-binding to regulatory elements. Single-cell technologies facilitate gene expression measurement for individual cells and cell-type identification, yet the connection between TF coordination and TG regulation of various cell types remains unclear. To address this, we developed a novel computational approach, Network Regression Embeddings (NetREm), to reveal cell-type TF-TF coordination activities for TG regulation. NetREm leverages network-constrained regularization using prior knowledge of direct and/or indirect PPIs among TFs to analyze single-cell gene expression data. We tested NetREm by simulation data and benchmark its performance in 4 real-world applications that have gold standard TF-TG networks available: mouse (mESCs) and simulated human (hESCs) embryonic stem (ESCs), human hematopoietic stem (HSCs), and mouse dendritic (mDCs) cells. Further, we showcased NetREm's ability to prioritize valid novel TF-TF coordination links in human Peripheral Blood Mononuclear cell (PBMC) sub-types. We applied NetREm to analyze various cell types in both central (CNS) and peripheral (PNS) nerve system (NS) (e.g. neuronal, glial, Schwann cells (SCs)) as well as in Alzheimer's disease (AD). Our findings uncover cell-type coordinating TFs and identify new TF-TG candidate links. We validated our top predictions using CUT&RUN (Cleavage Under Targets and Release Using Nuclease) and knockout loss-of-function expression data in rat/mouse models and compared results with additional functional genomic data, including expression quantitative trait loci (eQTL) and Genome-Wide Association Studies (GWAS) to link genetic variants (single nucleotide polymorphisms (SNPs)) to TF coordination. NetREm is open-source available on GitHub as a software package.

§ 3.1 Introduction

Transcription Factors (TFs) are proteins that work together to regulate (activate or repress) target gene (TG) expression in a coordination fashion, especially at the cell-type level (Lambert et al. 2018). TFs bind

to cognate DNA sequence-specific TF binding sites (TFBSs) on regulatory elements (e.g. enhancers and promoters) to mediate transcription of their respective TGs. Nonetheless, these gene regulatory mechanisms are multi-faceted; regulatory elements of TGs are formed by combinatorial interactions of multiple TFs within regulatory elements that form transcriptional regulatory modules (TRMs) (Guo and Gifford 2017) to ultimately govern transcription initiation (Nie et al. 2020). Most TFs cooperate with other TFs (rather than operate in isolation), working in concert to regulate gene expression utilizing mechanisms like co-binding or tethered-binding (Nie et al. 2020). TFs can be part of stable complexes (e.g. heterodimers (Ibarra et. al 2020)) or they can enhance binding affinity of other TFs to nearby TFBSs (synergistic activation) to regulate TGs (Ibarra et. al 2020; Zhao 2023). It is not uncommon for the regulation of one TG to necessitate interactions with 10-15 TFBSs (Bentsen et. al 2022). Despite the prevalence of such TF-TF coordination, the underlying intricacies of this phenomena are not yet fully comprehended (Ibarra et. al 2020). Further, a set of core TFs can determine cell-type-specific transcription profiles (Lee et al. 2012; D'Alessio et al. 2015). Indeed, TF binding grammar is complex (Bentsen et. al 2022), context-specific and cell-type-specific, depending on other proteins and chromatin structure around TFBSs. Since the diversity of TF binding is core for cell-type specificity (Lee et al. 2012), models of TG regulation by TFs (e.g. gene regulatory networks (GRNs)) must incorporate complex combinatorial analyses of TFs and their intricate interactions with one another.

Coordination among TFs regulates TG expression by modulating TF binding stability, localization, or post-translational modifications, affecting their regulatory function, activity, signaling. Contemporary research indicates a heightened adaptability among TFs in TG regulation, evidenced by phenomena like nucleosome-mediated cooperativity and the dynamic organization of TFBSs. These observations imply potential cooperativity even among disparate TFs lacking direct physical protein-protein contact (Mirny 2010; Badia-i-Mompel et al. 2023). Historically, classical models of TF cooperativity were predicated on direct protein-protein interactions (PPIs), which are physical contacts among proteins to handle various biological processes; for instance, to increase their binding affinity and motif specificity (Sönmezer et al. 2021), TFs may oligomerize to form transcriptional complexes to regulate TG expression

(Wang et al. 2023). However, it is now understood that TFs occupying the same regulatory region need not interact directly to jointly coordinate the transcriptional activation of a TG. That is, cooperativity among TFs can be via indirect PPIs as well (Rao et al. 2021). For example, the binding of a TF (e.g. pioneer TFs recognizing and binding to their TFBSs even in heterochromatin, closed/compact chromatin) can remodel local chromatin configurations or prompt DNA conformational transitions, such as nucleosomal DNA unwrapping or nucleosome eviction, thereby exposing TFBSs and rendering the DNA more amenable to subsequent binding by secondary or non-pioneer TFs (Mirny 2010; Rao et al. 2021; Sinha et al. 2023; Mayran and Drouin 2018). Additionally, through the tethered-binding mechanism involving coactivators and/or corepressors, TFs can exert regulatory influence over TGs via the direct or indirect cooperative recruitment of intermediary proteins, such as p300/CBP and the mediator complex (Spitz and Furlong 2012). In fact, recent studies suggest that most TF pairs that cooperate may form these DNA-mediated complexes (Ibarra et. al 2020). Even in the absence of direct PPIs, the coordinate effort of many TFs is indispensable for co-factor recruitment and the establishment of nucleosome-free regions, hallmark features of promoters and active enhancers (Rao et al. 2021), which characterize the distinct transcriptional profiles of specific cell-types. Antagonistic coordination, via sequestration and/or competition for TFBSs, generates cell-state heterogeneity by driving opposing effects on epigenetic programs and TG expression; some TFs bind DNA only in the presence of cooperating TFs and absence of antagonistic TFs (Hu et al. 2022; Berenson et al. 2023).

Recent single-cell data analyses (Mathys et. al 2019; Wang et. al 2018) show cell-type-specific expression dynamic patterns, implying that gene regulation is cell-type-specific. However, the mechanistic role and extent of TF associations with other TFs (e.g. cooperativity) to drive TG regulation across cell-types remain unclear (Nie et al. 2020; Ibarra et. al 2020; Karczewski et al. 2011; Hannenhalli and Levy 2002). For instance, neuronal and glial cells are important in nervous system development, function, repair and processes like myelination, synaptogenesis, neuroplasticity. Studies (e.g. (Joung et al. 2023)) have observed that overexpression of combinations of TFs can lead to massive alterations in GRNs, which may depend on the cell-type. Cell types play key roles in brain-related diseases. Microglia drive

neuroinflammation linked with Alzheimer's disease (AD) progression (Leng and Edison 2021) and excess synaptic pruning in Schizophrenia (SCZ) (Wang et al. 2019). Depression, Autism Spectrum Disorder (ASD), and SCZ may involve oligodendrocyte function (Maglorius Renkilaraj et al. 2017). Many TFs associated with SCZ and Bipolar Disorder (BD) have high expression in adult astrocytes (Pearl et al. 2019). Recent Attention-Deficit Hyperactivity Disorder (ADHD) studies identified causal risk genes highly expressed in fetal astrocytes, neurons, microglia (Fahira et al. 2019). Thus, understanding TF coordination, at the cell-type level, is crucial for dissecting the GRNs that govern processes fundamental to cognition, movement, behavior, and potential dysregulation in neurodegenerative diseases.

Despite extensive single-cell data analyses (Mathys et. al 2019; Wang et. al 2018), the extent of TF-TF coordination in regulating TGs across various cell types remains unclear (Nie et al. 2020; Ibarra et. al 2020; Karczewski et al. 2011; Hannenhalli and Levy 2002). Single-cell data is instrumental not only in identifying cell-type-specific biomarker genes but also in inferring cell-type-specific GRNs, which are crucial for understanding the development and maintenance of cellular identity and fates (Van de Sande et. al, 2020). GRN-inference tools often reverse engineer gene expression data to find coordinated patterns of expression and gene-gene interactions to predict the interplay between TFs and TGs, trying to uncover potentially meaningful biological signals from the noise (Wang et al. 2023; Campos et al. 2019; Zaborowski and Walther 2020). State-of-the-art (SOTA) tools create cell-type GRNs using various methods (e.g. co-expression, correlation r, information theory, differential equation, machine learning, multi-omics integration) but often overlook interdependent TF-TF PPI networks (PPINs) crucial for TG regulation. These tools presuppose that alterations in the expression of genes encoding TFs lead to subsequent changes in expression of their TGs (Zaborowski and Walther 2020) and infer potential cell-type-specific relationships among individual candidate TFs and their TGs that they help regulate (e.g. TF-TG regulatory links).

Differences among these tools is often attributed to their underlying assumptions applied to understand the nature of TF to TG regulatory dynamics (Nguyen et al. 2020) and methods (e.g. correlation) used to associate candidate TFs with their TGs (Kim et al. 2023). For example, the SCENIC (Aibar et. al

2017) pipeline utilizes co-expression analysis, such as GRNBoost (Moerman et al. 2019), on gene expression data to identify TGs for individual candidate transcription factors (TFs), focusing more on single TFs than on TF networks. SCODE (Matsumoto et al. 2017) uses ordinary differential equations on gene expression data to infer regulatory networks. PoLoBag utilizes polynomial lasso bagging for signed GRN inference (Roy et. al 2020), while TIGRESS (Haury et al. 2012) uses least angle regression (LARS) combined with stability selection. The BEELINE (Moerman et al. 2019) pipeline assesses 12 cell-type-specific GRN algorithms using a benchmark framework alongside gene expression data. However, GRN methods relying solely on single-omics data, typically single-cell gene expression, can miss vital molecular interactions and regulatory mechanisms influencing TF activity, such as chromatin accessibility, DNA methylation, histone modifications, and PPIs, including TF-TF interactions. Multi-omics integration tools (e.g. SCENIC+ (Bravo González-Blas et al. 2023), scGRNom (Jin et al. 2021), Signac (Stuart et. al 2021), Inferelator 3.0 (Skok Gibbs et al. 2022), CellOracle (Kamimoto et al. 2023)) which combine gene expression data with chromatin interactions, accessibility, and TF binding data, attempt to infer more comprehensive cell-type-specific GRN relationships among regulatory elements, TFs, and TGs.

Nonetheless, a key limitation of these GRN tools is they focus on TFs in isolation, often neglecting the network of critical interdependent TF-TF regulatory coordination (e.g. PPI links) that is essential for cell-type gene regulation. As a result, these tools may struggle to retain TF-TG pairs with weak or de-coupled relations (e.g. uncorrelated expression), potentially excluding fundamental and intricate aspects of gene regulatory mechanisms (Zaborowski and Walther 2020). For instance, studies have found that even uncorrelated expression between a given (TF, TG) pair may bury and obscure a true biological regulatory relationship; this TF may still be involved in regulating the TG, although perhaps via other complicated avenues (e.g. joint coordination with other TFs)(Zaborowski and Walther 2020). In fact, cooperation and competition among TFs (i.e. TF binding process) has been shown to increase the noise in the gene expression profiles of TGs (Parab et al. 2022). These coordinating TF predictors are often co-expressed (significant magnitude of correlation with each other and high multicollinearity), impacting TG expression levels through synergistic or antagonistic coordination (Parab et al. 2022); however, in the

absence of additional context (e.g. PPIs for TFs), these GRN inference tools may potentially select independent TFs and/or remove some of these co-expressed TF predictors that are truly causal for the TG expression (Nicodemus and Malley 2009). In previous studies, PPIs have been used to infer synergistic binding of cooperative TFs (Nagamine et al. 2005) and classify the nature of interactions among TFs (Perna et al. 2020). Thus, there is a need for these GRN-inference tools to incorporate different levels of gene expression regulation, including PPIs (Zaborowski and Walther 2020). The BGRMI tool solely uses PPIs to learn relevant terms for TF heterodimer complexes to add to the model, ignoring the inherent PPI network structure (including indirect associations) and corresponding network weights (Iglesias-Martinez et al. 2016). While SCINET (Li and Li 2008) reconstructs cell-type-specific interactomes by integrating a reference interactome with gene expression data, it does not explicitly reveal cell-type gene regulation. TF-Cluster (Nie et al. 2011) identifies functionally coordinated TFs involved in biological processes but does not focus on TG regulation and is based on coexpression analysis (and does not use any existing prior knowledge). RTNduals (Chagas et al. 2019) predicts TF co-regulatory behavior solely from expression data, which may not always yield outputs, and does not return a predicted TF-TG regulatory network.

Previous studies have leveraged network-regularized (e.g. graph-regularized, network Lasso) regression models to identify disease-associated genes and gene networks, incorporating existing biological information and metadata as prior knowledge (Li and Li 2008; Wang et al. 2015; Dirmeier et al. 2018; Li and Li 2010; Kim et al. 2013). This biological knowledge guides and constrains the regression problem, helping improve the biological relevance of the final inferred regression model. Nonetheless, these models have not been tailored to elucidate TF coordination in gene regulation. Further, there is a pressing need for these network-regularized regression models to evolve, enabling them to learn from regression data and prior networks while also creating robust latent embedding representations from these inputs. Several studies (e.g. (Chu et al. 2023; Gharavi et al. 2021; Choy et al. 2019)) have showcased the power of embeddings for downstream analysis, their efficacy at extracting significant relationships and facilitating new insights in biology and in other contexts. To bridge these gaps, we developed NetREm

(Network Regression Embeddings), a novel computational framework designed to infer cell-type TF coordination activities for gene regulation.

Building on established network-regularized regression techniques, NetREm integrates multimodal data (e.g. TF binding profiles, direct/indirect TF-TF PPIs, derived TF-TF colocalization, gene expression, chromatin interaction, scATAC-seq epigenomic markers) capturing intricate aspects of TG regulation. It constructs robust predictive models for TF-TG regulation (complementary GRNs) as well as TF-TF coordination. A distinct feature of NetREm is its innovative ability to generate network regression embeddings, which identify and quantify coordination among cell-type TFs for co-regulating individual TGs. Public databases like STRINGdb provide direct and/or indirect, organism-specific, global, undirected, cell-type-agnostic PPIs for >12k organisms (Szklarczyk et al. 2023). Despite PPINs having some incorrect PPIs (False Positives (FPs)) and being largely incomplete (Kotlyar et. al 2022), it is helpful to integrate TF-TF PPINs as prior information (McCalla et al. 2023; Li and Jackson 2015; Ghanbari et al. 2015; Imoto et al. 2003; Mukherjee and Speed 2008). This improves NetREm's predicted GRNs from expression data. NetREm also addresses the need for not only cell-type-specific annotation of known PPIs, but also discovery of new PPIs. It provides much-needed insights into TRMs, reveals cell-type- and disease-specific TF-TF PPINs, and aids interactome studies in uncovering disease gene properties and differential PPIN rewiring (Göös et al. 2022; Sevimoglu and Arga 2014). That is, NetREm can reveal TRMs comprising co-associated TFs and can prioritize TGs and coordination among TFs that significantly vary across different cell types or disease states. NetREm may help uncover and prioritize TF-TF cell-type-specific and disease-specific interactions that are subnetworks of the original PPI. This may aid studies that use the human interactome to uncover network properties of disease genes and differential network rewiring that is context-specific (Sevimoglu and Arga 2014). Further, this may enhance our understanding of how direct and/or indirect PPIs among TFs play roles in transcription regulation, which is currently poorly known (Göös et al. 2022). As demonstrations, we applied NetREm to simulation data and various cell types in both central and peripheral nerve systems (PNS) such as myelinating (mSCs) and non-myelinating

(nmSCs) SCs, as well as in Alzheimer's disease (AD) and control states in various neuronal and glial cells. However, NetREm is an open-source tool for general-purpose use.

§ 3.2 Methods and Materials

3.2.1 NetREm Methodology

Overall, NetREm provides a holistic approach to understanding cell-type-specific gene regulatory mechanisms and enriches our understanding of core cell-type interactions (direct and indirect) among TFs that are involved in these processes. Please see **§B.1** (SaniyaKhullar 2024) for more details on the methods for NetREm.

Integrating multimodal data and networks in NetREm workflow

Preliminary definitions: Proteins (e.g. TFs), and TGs (*italics*) are represented by HGNC symbols. TF gene expression is a proxy for TF protein abundance, assuming high gene expression (quantifies mRNA abundance) translates to high protein expression. NetREm can be applied to both single-cell and bulk expression data (sample by gene). Bulk data represents pooled collections of cell lines or tissues, yielding averaged expression profiles, as seen in patients.

We start with single-cell gene expression data for M samples (individual cells) and $\mathcal{T}$ genes in a cell-type; scRNA-seq data is typically high-dimensional ($M \ll \mathcal{T}$), sparse, non-negative. We focus on G TGs, (i.e. $\{\mathrm{TG}_k\}_{k=1}^{G}$), with expression profiles $\{y_k\}_{k=1}^{G}$, respectively, and $\mathcal{N}$ potential cell-type TFs, where $G \leq \mathcal{T}$, $\mathcal{N} < \mathcal{T}$. If all TFs are master regulators: $\mathcal{T} = G + \mathcal{N}$; else (some TGs are TFs): $\mathcal{T} < G + \mathcal{N}$. We ensure that a given TG is not its own candidate TF.

NetREm can use prior GRN information that identifies $\{N_k\}_{k=1}^{G}$ respective promising candidate TFs for the set of G TGs where $N_k \leq \mathcal{N}$ and varies based on the optimal prior GRN TFs selected for the given TG_k. When prior GRN information is absent, TG_k (and other $G - 1$ TGs) will have the same fixed $N_k = \mathcal{N}$ candidate TFs; it is to be understood that if TG_k is also a TF, its self-TF is excluded, so it has $N_k = \mathcal{N} - 1$ candidate TFs. These steps below are repeated for each of the G TGs. For simplicity, we

explain the pipeline for predicting expression of a single TG. We use N to represent the # of its candidate TFs for this TG, and y for its true expression.

Optional (recommended) prior cell-type GRN information: Constructing candidate links from TFs to TGs may improve the quality of NetREm's solution via initial feature selection of N biologically meaningful TFs tailored for TG, where $N < \mathcal{N}$. Using diverse information on expression regulation guides the regression (Zaborowski and Walther 2020). We integrated prior GRNs (initial TF-RE-TG links) from multi-omics data in various applications related to Schwann cell sub-types (application 6) and to Alzheimer's disease (AD) versus controls in 8 neuronal/glial cell-types (application 7). When no prior GRN is used: if TG is a TF, it has $N = \mathcal{N} - 1$ candidate TFs, otherwise it has all $\mathcal{N}$ TFs as candidates.

PPI network (PPIN, network prior): Our comprehensive, weighted, undirected PPIN $\mathbb{W}$ illuminates biological interactions among proteins (network nodes) with strong functional association evidence in the organism. It inherently captures direct (e.g. complex formation, transient interactions) and/or indirect (e.g. participate in shared processes or bind by intermediate hidden partners (De Las Rivas and Fontanillo 2012)) PPIs. Alas, PPINs attribute weight $w > 0$ to all PPIs, even to those with antagonistic, competing roles in paths (Szklarczyk et al. 2023); each edge w accounts for uncertainty in $\mathbb{W}$ and is proportional to the probability that the 2 connected nodes interact (i.e. their integrative functional essentiality in $\mathbb{W}$) (Li and Liu 2022). For the TG, we subset $\mathbb{W}$ to obtain $W^{(0)}$ that captures known TF-TF PPIs, reflects potential structure/relation background information among TG's N candidate TF proteins, and is symmetric (i.e. $\left(W^{(0)}\right)^T = W^{(0)}$). Higher w_{ij} denotes more confidence that TF_i and TF_j partner, directly and/or indirectly, in processes like regulating DNA chromatin loops of interacting REs for TG regulation (Wang et al. 2021a). Here, $w_{min} = \min\{w_{ij} \in W^{(0)}\}$ is the smallest confidence for known direct/indirect TF-TF links among N TFs. To enable NetREm to consider candidate TFs missing from $W^{(0)}$, we add artificial weight $0 < \eta < w_{min}$ for missing pairwise edges; these novel edges may not exist (i.e. TFs truly do not coordinate: True Negatives) or are yet to be discovered (FNs). We obtain our final, fully-connected, TG-specific, TF-TF input PPIN: $W \in \mathbb{R}^{N \times N}$ where $w_{ij} = w_{ji} > 0$. W has

$\frac{N(N-1)}{2}$ unitless, global PPIs with $w_{ij} > 0$ equal to: $w_{ij}^{(0)}$ for known and η for artificial TF-TF links. We do not consider self-loops; instead, we set $W_{ii} = \frac{d_i}{N-1}$ where TF_i's degree (connectivity) with other $N-1$ TFs is $d_i = \sum_{k \neq i} w_{ik} > 0$.

$$W = \begin{bmatrix} \frac{\sum_{j=2,\ i\neq j}^{N} w_{1j}}{N-1} & w_{12} & \cdots & w_{1N} \\ w_{21} & \frac{\sum_{j=1,\ i\neq j}^{N} w_{2j}}{N-1} & \cdots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & \frac{\sum_{j=1, i\neq j}^{N} w_{Nj}}{N-1} \end{bmatrix}_{N \times N} = \begin{bmatrix} \frac{d_1}{N-1} & w_{12} & \cdots & w_{1N} \\ w_{21} & \frac{d_2}{N-1} & \cdots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & \frac{d_N}{N-1} \end{bmatrix}_{N \times N}.$$

NetREm integrates input data: For the M samples, the $X \in \mathbb{R}^{M \times N}$ matrix contains expression data for N predictors, while $y^{(0)} \in \mathbb{R}^M$ is the expression vector for TG. We standardize each column of X ($X_{ij} \leftarrow \frac{X_{ij}-\mu_j}{\sigma_j}$) and y ($y_i \leftarrow \frac{y_i-\mu_y}{\sigma_y}$) by respective means μ_j, μ_y and standard deviations σ_i, σ_y, making them unitless. Then, each TF in X has $\bar{\mu_j} \approx 0$ and $\sigma_j \approx 1$; $\mu_y \approx 0$, $\sigma_y \approx 1$. Pairwise r are preserved. If $M < N$, X suffers the curse of dimensionality. With technological advances and the advent of large single-cell sequencing studies, we anticipate a boost in M so $M \gg N$ will soon be the norm (Cuomo et al. 2023), especially since $N \ll \mathcal{T}$ (relatively few genes are transcribed and translated to proteins that are TFs). NetREm identifies which of N TFs can predict TG expression y, considering PPIs among TFs. NetREm can comprehensively integrate multi-omics data and PPINs to discover key TFs and TF-TF coordination events for TG regulation in a cell-type-specific manner.

Step 1: Network regularized regression

Given gene expression data with M samples (rows) and N features (columns) represented as $X \in \mathbb{R}^{M \times N}$ (e.g. N candidate TF expression values for each M) and response $y \in \mathbb{R}^M$ (e.g. TG expression for each M), we want to learn a linear predictor $y \approx Xc^*$ for *TG.* By imposing prior TF-TF PPIN information as a regularization term, we develop a functional map highlighting connectivity patterns among N TFs for TG; this typically steers NetREm to favor groups of TFs with shared PPIN connectivity over isolated TFs, thereby enhancing its ability to capture biologically relevant PPIs (Li and Liu 2022). Coefficients $c^* \in \mathbb{R}^N$ represent the importance of TFs for regulating TG and are found by optimizing the following problem with objective function:

$$c^* = \underset{c}{\operatorname{argmin}} f(c) = \frac{1}{2M}\left||y - Xc|\right|^2 + \alpha\left||c|\right|_1 + \frac{\beta}{2}\sum_{i=1}^{N}\sum_{j=i}^{N} w_{ij}\left(\frac{c_i}{\sqrt{d_i}} - \frac{c_j}{\sqrt{d_j}}\right)^2 \qquad Eq\ (1)$$

Our 3 terms in equation $Eq\ (1)$ are unitless and compatible for addition:

1 (data-fitting): ensures Xc is close to y, and $\frac{1}{2M}$ is a normalization factor to make it invariant to sample size M. **2** (sparsity-prior): favors a sparse solution (small # of non-zero c^*), helping simplify the model and boost reliability. **3** (network-prior): penalizes differences between c^* of connected TF nodes, normalized by their respective network centrality d and adjusted for their global (organism-based) PPI weights with other candidate TFs in TG-specific input W. Inspired by (Li and Li 2008), this approach allows for a more equitable representation of TFs, irrespective of their d; this network-oriented variant of Ridge L_2 penalty $\sum_{i=1}^{N} c_i^2$, promotes topology-aware c^* shrinkage and smoothing for neighboring TF_i and TF_j, with probability proportional to w_{ij}. It underscores the principle that strongly connected TFs likely perform shared functions, even if their influence on TG expression (c^* signs) differs. This recognizes the community structure in existing PPINs that groups proteins with similar biological roles with $w > 0$, not distinguishing between cooperative (+) and antagonistic (-) PPIs (Padi and Quackenbush 2015; Szklarczyk et al. 2023). NetREm leverages this refined understanding, offering a comprehensive perspective on the interplay of TFs in W.

We can tune 2 hyperparameter knobs: network-constrained prior, $\beta > 0$, decides the strength of the PPIN regularization penalty (applied 1st: higher β guides NetREm to prioritize TFs with strong PPIs); sparsity prior, $\alpha \geq 0$, impacts $\mathcal{L}_1$ penalty (applied 2nd). NetREm, a PPIN-aware adaptation of ElasticNet, performs automatic variable selection based on expression and PPIN data, grouping and selecting strongly-connected TFs (emphasizing known TF-TF PPI subnetworks) in a spirit akin to ElasticNet. If $M \ll N$, ElasticNet and NetREm may still select $\leq N$ TFs as final; this addresses limitations of Lasso regression that may indiscriminately select only 1 TF from a group of highly correlated TFs and only $\leq M$ TFs if $M \ll N$ (Zou and Hastie 2005; Li and Li 2008).

Step 2: Gene embeddings from network regression

Our novel method transforms the original problem into a Lasso regression problem in a new space with cell-type-specific TF-TF interactions. 1st, we represent the network-prior term in a more compact matrix-vector form as: $\frac{\beta}{2}\sum_{i=1}^{N}\sum_{j=i}^{N} w_{ij}\left(\frac{c_i}{\sqrt{d_i}}-\frac{c_j}{\sqrt{d_j}}\right)^2 = \frac{\beta}{2}c^T Ac$, where $A = D^T(W \odot V)D =$

$$\begin{bmatrix} 1 & -w_{12}/\sqrt{d_1 d_2} & \cdots & -w_{1N}/\sqrt{d_1 d_N} \\ -w_{21}/\sqrt{d_1 d_2} & 1 & \ddots & -w_{2N}/\sqrt{d_2 d_N} \\ \vdots & \vdots & & \vdots \\ -w_{N1}/\sqrt{d_1 d_N} & -w_{N2}/\sqrt{d_2 d_N} & \cdots & 1 \end{bmatrix}.$$ A, D, V, W are all symmetric $\mathbb{R}^{N\times N}$ matrices.

We define $V = N \cdot I - 11^T$ where $1 \in \mathbb{R}^{N\times 1}$ is a N dimensional all 1 column vector, and $W \odot V$ is element wise ($\odot$) multiplication of W and V (i.e. Hadamard product). $D = diag\left(1/\sqrt{d}\right)$ is a diagonal matrix with main diagonal degree-based elements $1/\sqrt{d_i}$ and off-diagonals 0: $D =$

$$\begin{bmatrix} \frac{1}{\sqrt{d_1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{d_2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{d_N}} \end{bmatrix}_{N\times N} = D^T$$ where $d_i = \sum_{j=1,\ j\neq i}^{N} w_{ij}$. V is invariant to W, D, and input TFs and TGs;

instead V only depends on N (# of candidate TFs) to obtain its constant values. That is, $V =$

$$\begin{bmatrix} N-1 & -1 & \cdots & -1 \\ -1 & N-1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & N-1 \end{bmatrix}_{N\times N}$$ and $W \odot V = \begin{bmatrix} d_1 & -w_{12} & \cdots & -w_{1N} \\ -w_{21} & d_2 & \cdots & -w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ -w_{N1} & -w_{N2} & \cdots & d_N \end{bmatrix}_{N\times N}$.

Given that squared term $\left(\frac{c_i}{\sqrt{d_i}}-\frac{c_j}{\sqrt{d_j}}\right)^2 \geq 0$ and $w_{ij} \geq \eta > 0$ and $\beta > 0$, the quadratic form $c^T Ac \geq 0$ for any vector of optimal c^* and $A = [a_{ij}] \in \mathbb{R}^{N\times N}$ is therefore symmetric and positive semi-definite. Here, $a_{ij} = a_{ji} < 0$ for $i \neq j$ and main diagonal $a_{ii} = 1$ for $i = 1, \ldots, N$. A scales each entry $(W \circ V)_{ij}$ by $1/\sqrt{d_i d_j}$ so $A_{ij} = (W \odot V)_{ij}/\sqrt{d_i d_j}$. Here, A captures connectivity and interaction strengths in W in a form suitable for regularized regression. As A is based on a fully-connected PPIN (i.e. there are no completely isolated subsets of TF nodes), it is a normalized variant of normalized graph Laplacian matrix L, specifically tailored for context where w and d are crucial. Off-diagonals are $-1 \leq A_{ij} = -w_{ij}/\sqrt{d_i d_j} < 0$ where $i \neq j$, since $d_i = \sum_{k\neq i} w_{ik} \geq w_{ij}$ and $d_j = \sum_{k\neq j} w_{jk} \geq w_{ji}$, so $\sqrt{d_i d_j} \geq w_{ij}$;

negative values indicate respective penalty for dissimilarity between connected nodes since the regularization term aims to minimize differences in characteristics (modeled by $|c^*|$) amongst TFs by regularizing and smoothing c^* across input PPIN based on connectivity. Using this matrix-vector representation, we reformulate $f(c)$ in $Eq\ (1)$ as

$$f(c) = \frac{1}{2M}\left|\left|y - Xc\right|\right|^2 + \alpha\left|\left|c\right|\right|_1 + \frac{\beta}{2}c^T Ac$$

We note that $\left|\left|y - Xc\right|\right|^2 = (y - Xc)^T(y - Xc) = y^T y - 2y^T Xc + c^T X^T Xc$. Thus,

$$f(c) = \frac{1}{2M}(y^T y - 2y^T Xc + c^T X^T Xc) + \alpha\left|\left|c\right|\right|_1 + \frac{\beta}{2}c^T Ac$$

$$= \frac{1}{2M}c^T(X^T X + \beta MA)c\ -\ \frac{1}{M}y^T Xc +\ \alpha\left|\left|c\right|\right|_1 +\ \frac{1}{2M}y^T y$$

We set $E = \frac{X^T X}{M} + \beta A \in \mathbb{R}^{N \times N}$, which we derive by dividing $X^T X + \beta MA$ by M. E is symmetric and positive semi-definite since A and Gram matrix $X^T X \in \mathbb{R}^{N \times N}$ are symmetric and positive semi-definite and $\beta, M, N > 0$. Here, $(X^T X)_{ii} = M$, reflecting sum of squared values of each TF, indicative of variance ($\sigma^2 = 1$) scaled by M. Off-diagonals $(X^T X)_{ij}$ (for $i \neq j$) represent sums of products of pairs of different TFs so $\left|(X^T X)_{ij}\right| \leq M$ since $\frac{(X^T X)_{ij}}{M}$ represents: $\left|cor\left(TF_i, TF_j\right)\right| \leq 1$. Thus, this 1st gene expression-based term (i.e. cell-type-specific data term) for E is: $\frac{X^T X}{M}$, representing correlation r among TFs, has max value 1, is the covariance matrix of X scaled by $\frac{1}{M}$. The 2nd term (i.e. cell-type-independent PPI term): βA is β for main-diagonals and has off-diagonals $-\beta \leq \beta A_{ij} < 0$ (for $i \neq j$). As region-specific PPINs become available for the given organism, they may be used as input to NetREm in place of the current, global, organism-specific PPINs to improve the cell-type-specificity of the results. Thus, $E_{ij} = r\left(x_i, x_j\right) - \beta\frac{w_{ij}}{\sqrt{d_i d_j}}$, which balances r with PPIN knowledge (for a given β value) and information $\left(\frac{w_{ij}}{\sqrt{d_i d_j}}\right)$.

$$f(c) = \frac{1}{2M}c^T(X^T X + M\beta A)c\ -\ \frac{1}{M}y^T Xc +\ \alpha\left|\left|c\right|\right|_1 +\ \frac{1}{2M}y^T y$$

$$= \frac{1}{2N}c^T\tilde{X}^T\tilde{X}c\ -\ \frac{1}{N}\tilde{y}^T\tilde{X}c + \alpha\left|\left|c\right|\right|_1 + \frac{1}{2M}y^T y$$

$$= \frac{1}{2N}\left|\left|\tilde{y} - \tilde{X}c\right|\right|^2 + \alpha\left|\left|c\right|\right|_1 + \frac{1}{2M}y^T y - \frac{1}{2N}\tilde{y}^T\tilde{y}, \qquad Eq\ (2)$$

where $\tilde{X} \in \mathbb{R}^{N \times N}$ and is symmetric (i.e. $\tilde{X} = \tilde{X}^T$) and $\tilde{y} \in \mathbb{R}^N$ satisfies:

$$\frac{1}{N}\tilde{X}^T\tilde{X} = \frac{1}{M}X^T X + \beta A \qquad Eq\ (3a)$$

$$\frac{1}{N}\tilde{y}^T\tilde{X} = \frac{1}{M}y^T X \qquad Eq\ (3b)$$

Finally, we reformulate $Eq\ (2)$ as a conventional Lasso problem (by omitting constant term $\frac{1}{2M}y^T y - \frac{1}{2N}\tilde{y}^T\tilde{y}$) that we solve using existing standard Lasso or LassoCV (Pedregosa et al. 2011) solvers:

$$c^* = \underset{c}{\operatorname{argmin}}\, \tilde{f}(c) = \frac{1}{2N}\left|\left|\tilde{y} - \tilde{X}c\right|\right|^2 + \alpha\left|\left|c\right|\right|_1 \quad Eq\ (4)$$

To compute $\tilde{X}$ and $\tilde{y}$ we perform a Singular Value Decomposition (SVD) on E expressed as: $E = U\Sigma U^T$. Here $U \in \mathbb{R}^{N \times N}$ is the matrix of the left singular vectors of E and $\Sigma \in \mathbb{R}^{N \times N}$ is a diagonal matrix of singular values $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$ of E. All N values in $\mathcal{S}$ are non-negative and convey info regarding strength or importance of each corresponding dimension ($s_{\max} = \max(\mathcal{S})$ and $s_{\min} = \min(\mathcal{S})$). Then, $E = U\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}U^T = \left(\Sigma^{\frac{1}{2}}U^T\right)^T\left(\Sigma^{\frac{1}{2}}U^T\right)$. Based on $Eq\ (3a)$, $E = \frac{1}{N}\tilde{X}^T\tilde{X}$. Then, $\tilde{X}^T\tilde{X} = NE$. Hence, $\left(\sqrt{N}\Sigma^{\frac{1}{2}}U^T\right)^T\left(\sqrt{N}\Sigma^{\frac{1}{2}}U^T\right) = \tilde{X}^T\tilde{X}$. For improved stability, we set small singular values (e.g., $s_i < 10^{-6}s_{\max}$) to 0, resulting in truncated Σ_{trunc}, and similarly, we adjust inverse ${\Sigma_{\text{trunc}}}^{-1}$ setting inverse elements corresponding to small singular values to 0. By substituting ${\Sigma_{\text{trunc}}}^{-\frac{1}{2}}$ in place of $\Sigma^{-\frac{1}{2}}$, we effectively use truncated SVD, enhancing NetREm's robustness by excluding contributions from small singular values. Thus, X and y are transformed to a new latent space of gene expression embeddings that incorporate PPIN information: $\tilde{X} \in \mathbb{R}^{N \times N}$ and $\tilde{y} \in \mathbb{R}^N$, respectively, in $Eq\ (4)$:

$$\tilde{X} = \sqrt{N}{\Sigma_{\text{trunc}}}^{\frac{1}{2}}U^T \text{ where } \tilde{X} = \begin{bmatrix} | & | & \cdots & | \\ \widetilde{X_1} & \widetilde{X_2} & \ldots & \widetilde{X_N} \\ | & | & \cdots & | \end{bmatrix}_{N \times N}$$

$$\tilde{y} = \frac{\sqrt{N}}{M}{\Sigma_{\text{trunc}}}^{-\frac{1}{2}}U^T X^T y$$

When $\beta = 0$, the transformation yields X's principal components (PCs) via SVD on $X^T X$, a trivial case without PPIN information; however, since we require $\beta > 0$, $\tilde{X}$ not only reflects its PCs but also includes

PPIN structure, with βA added to $\frac{1}{M}X^TX$, creating an "embedding". This comprehensive approach captures both data patterns and PPI relations. The higher β is, the greater the contribution of PPIN relations will be towards $\tilde{X}$ and $\tilde{y}$, which encapsulates expression relations and PPIN information. We perform Lasso regression to solve $Eq\ (4)$ for $\tilde{X}$ and $\tilde{y}$, determining optimal c^*. **§B.1** (SaniyaKhullar 2024) provides more details.

Output 1: Identification of potential novel cell-type TFs in TF-TG regulatory network

TF-TG regulatory network for TG:

Several TFs regulate transcriptional activity of TG in a cell-type at a certain time (Nie et al. 2011). Solving the network-regularized regression problem produces a vector of Lasso $c^* \in \mathbb{R}^N$ for N TFs predicting true TG expression y. We focus on $c^* \neq 0$, which represents N^* final TFs for TG out of N candidates, where $0 < N^* \leq N$. NetREm constructs a comprehensive directed TF-TG regulatory network (complementary GRN) of N^* edges, weighted by c^*. Here $|c_i^*|$, indicates the strength of the TF_i-TG link, measuring TF_i's relative importance in regulating TG. TF_i with $c_i^* > 0$ may activate TG and TF_j with $c_j^* < 0$ may repress TG transcription and subsequent expression. This tug-of-war between activators and repressors orchestrates TG regulation. Biological complexity enables certain TFs to have dual-function roles, alternating between activation and repression depending on context and signals (Skok Gibbs et al. 2022; Boyle and Després 2010). In fact, the role of the TF in regulating TG expression (i.e. activator versus repressor role) may be governed by a spatial grammar (e.g. precise position of TF relative to Transcription Start Site (TSS) of TG, spatial configuration of TFBSs)(Duttke et al. 2024). Given the competitive nature of TFBS binding, $N - N^*$ discarded TFs may lose to some of the N^* TFs (i.e. antagonistic relation), but we do not speculate on this. Overall, NetREm unearths novel cell-type-specific, coordinating TFs involved in TG regulation, providing a more nuanced view of TF-TG interactions.

We evaluated our performance in training and testing expression data, comparing predicted $\widehat{\tilde{y}} \in \mathbb{R}^{N \times 1}$ to actual $\tilde{y}$ using metrics like Mean Square Error (MSE) $= \frac{1}{N}\sum_{v=1}^{N}\left(\widehat{\tilde{y}_v} - \tilde{y}\right)^2$. To achieve more accurate regulatory links, we integrated multiomics data like TF-DNA-binding (Dibaeinia and Sinha 2020)

in applications 6-7 to predict prior GRNs. We identified regulatory elements (REs) for TG and determined TFs likely to bind directly to or associate indirectly with TFBSs on these REs. We input N TG-specific candidate TFs to NetREm for TG. Then, we overlaid NetREm's N^* TF-TG regulatory links for TG with this prior GRN (initial TF–RE –TG links for N TFs for TG). This helped us annotate our links with epigenomic information on REs. Ultimately, we isolated highly-confident final TF-RE-TG links for our final N^* TFs.

Cell-type TF-TG regulatory network:

We applied NetREm, iteratively, to each of the G TGs and weave together individual TF-TG links (details: **§B.1** (SaniyaKhullar 2024)). We may narrow down links by retaining TFs with $|c^*| > c_{min}$ (min threshold, default: 0) and TGs meeting specific criteria (e.g. $MSE_{TG} < MSE_{max}$). Our cell-type-specific complementary GRN relates TFs to TGs they regulate, helping explain how cell-types establish and maintain cellular identity. We may annotate/validate this network by identifying eSNPs impacting TF binding with eQTL links to altered TG expression (Coetzee et al. 2015); when prior GRNs are used, we ensure SNPs fall in the same REs where TFs are predicted to bind, linking them to TG regulation.

Output 2: Cell-type-specific TF-TF coordination (direct/indirect TF-TF interactions)

TG-specific cell-type coordination B:

NetREm helps fulfill the need for cell-type-specific proteome analysis by which proteins interact to carry out processes like TG regulation. Existing PPINs aggregate direct and/or indirect PPIs in an organism. This broad approach has limitations, as not all proteins are expressed in every cell or tissue type, and some may be aberrant in diseases (Padi and Quackenbush 2015). To overcome this, there are efforts to annotate global PPIs at various levels, including tissue-specific protein expression, cell-line-specific links, phenotype-based studies (e.g. CPPID) (Federico and Monti 2020). However, co-expression of TFs need not imply they interact in specific cell or tissue types (Gonzalez-Teran et al. 2022). Further, PPINs do not distinguish between cooperation and antagonism. This underscores the need for NetREm's 2nd output that predicts how TFs coordinate to regulate TG in the cell-type.

NetREm's 2nd output is a weighted and signed TG-specific TF-TF coordination network given by an adjacency matrix of coordination scores B, if $N^* \geq 2$ final TFs for TG. These scores are a function of both embeddings $\tilde{X}$ and c^*: $B = f(\tilde{X}, c^*)$. In our framework, $B_{ij} > 0$ suggests cooperativity (e.g. co-binding, pioneer-settler TF relations) and $B_{ij} < 0$ indicates antagonism (e.g. sequestration) between TF_i and TF_j for co-regulating TG.

We used c^* to predict the nature of interactions among N^* TFs for TG in a symmetric matrix $C \in \mathbb{R}^{N \times N}$. Here, $C_{ij} = sign(c_i, c_j) = \{1 \text{ if } c_i^* c_j^* > 0;\ -1 \text{ if } c_i^* c_j^* < 0;\ 0 \text{ otherwise}\}$ and $C_{ii} = 0$. RTNduals assesses coordinated behavior of 2 TFs by analyzing correlation distributions between them and their shared TGs; building on approaches like these, we use C to deduce relative coordination relations among TFs for TG, acknowledging TFs may exhibit antagonistic or cooperative interactions depending on TG and context. If $C_{ij} > 0$, both TFs likely cooperate, aiming to either upregulate or downregulate TG expression in unison; their combined synergistic net effect on the TG is stronger than their individual effects. Conversely, if $C_{ij} < 0$, both likely act antagonistically, with conflicting influences on TG expression; this activator-repressor antagonism weakens their combined effect compared to their individual impacts, potentially due to partially canceling each other's activities (Berenson et al. 2023). When $C_{ij} = 0$ (and $i \neq j$), at least 1 of the 2 TFs is not a final TF and we cannot ascertain their potential nature of interaction.

Earlier, we set $E = \frac{X^T X}{M} + \beta A$. The 1st term $\frac{X^T X}{M}$ represents the original normalized inner product space from column vectors $x_1, x_2, \ldots, x_N$ of X, where $x_i \in R^M$ represents TF_i's standardized expression levels across M cells in cell type. The 2nd term $A = D^T(W \odot V)D$ purely depends on TF-TF PPIN strengths $w > 0$ and can be retrieved from public databases. By $Eq\ (3a)$, $E = \frac{X^T X}{M} + \beta A = \frac{1}{N}\tilde{X}^T\tilde{X}$; we thus transform X and network-prior PPIN data to $\tilde{X}$ embedding data. This yields a new normalized inner-product space $\frac{1}{N}\tilde{X}^T\tilde{X}$ that helps depict and encode an aspect of cell-type TF-TF coordination scores for regulating TG. Since N is a scalar, we use $\left|\tilde{X}^T\tilde{X}\right| \in \mathbb{R}^{N \times N}$. For each $TF_i - TF_j$ pair, we divide $\left|\widetilde{X_i}^T \widetilde{X_j}\right|$,

which is proportional to the extent of their potential coordination, by $\left|\left|\widetilde{X_i}\right|\right| \cdot \left|\left|\widetilde{X_j}\right|\right|$ to scale it. $\left|\left|\widetilde{X_i}\right|\right| = \sqrt{\sum_{z=1}^{M}(\widetilde{x_{iz}})^2} > 0$ is the Euclidean norm of TF_i's embedding $\widetilde{X_i}$. This essentially is their cosine similarity (cos) magnitude: $\left|\cos\left(\widetilde{X_i}, \widetilde{X_j}\right)\right| = \frac{\left|\widetilde{X_i}^T \widetilde{X_j}\right|}{\left|\left|\widetilde{X_i}\right|\right| \cdot \left|\left|\widetilde{X_j}\right|\right|} \leq 1$. To learn coordination scores, we use coefficient-aware-cos metric: $B_{ij}^{(0)} = \left|\cos\left(\widetilde{X_i}, \widetilde{X_j}\right)\right| \odot C$ for $i \neq j$ and $B_{ii}^{(0)} = 0$. We apply max absolute value scaling $\frac{B_{ij}^{(0)}}{\max(|B|)}$ where $\max(|B|)$ is the max magnitude of $\ell_N = \frac{N(N-1)}{2}$ scores. Our TG-specific TF-TF coordination $B \in \mathbb{R}^{N \times N}$ has: $B_{ij} = \frac{100 B_{ij}^{(0)}}{\max(|B|)}$ where $-100 \leq B_{ij} \leq 100$ if TF_i and TF_j are among N^* TFs (i.e. $c_i^* c_j^* \neq 0$) where $i \neq j$; else B_{ij} is 0 for all remaining cases. Of ℓ_N scores, $\ell_{N^*} = \frac{N^*(N^*-1)}{2}$ are $\neq 0$.

TFs with higher $|B|$ have stronger coordination for co-regulating TG. NetREm predicts B for known TF-TF PPIs (pairs with $w_{ij} > \eta$), uncovering meaningful, cell-type-specific PPI subnetworks. These documented PPIs have established partnership for orchestrating biological processes. It also predicts B for novel, artificial PPI links ($w_{ij} = \eta$), flagging (high $|B|$) promising FN TF-TF links for follow-up investigation. TF-TF coordination can be direct (e.g. form complexes, tethering) or indirect (e.g. TFs may not interact physically but can modify local chromatin environments, facilitating binding of other TFs) (Srivastava and Mahony 2020). Typically, direct PPIs rely on genome-wide data, indirect (e.g. guilt-by-association) PPIs use genetic interaction data (Wang et al. 2009). While studies predict negatomes (proteins unlikely to interact physically) (Jha et al. 2022), they offer limited insights on indirect PPIs.

Cell-type coordination $\bar{B}$:

Functionally-coordinated TFs help regulate sets of TGs at a given time point (Nie et al. 2011) through direct and/or indirect interactions with one another. In reality, TF_i and TF_j may only co-regulate a subset of the G TGs in the cell-type, if at all. Nonetheless, their coordination behavior is considered across all G TGs to implicitly adjust for this; that is, that the learned cell-type TF-TF coordination score $\bar{B}_{ij}$ will reflect better their coregulatory behavior for the cell-type.

To create a weighted and signed cell-type-specific coordination network among $\mathcal{N}$ TFs, $\bar{B}$, we first aggregate individual B across G TGs. Each TG (i.e. TG_k) has expression (y_k), N_k candidate TFs (where $N_k \leq \mathcal{N}$), and coordination score matrix $B^k \in \mathbb{R}^{N_k \times N_k}$, where $k = 1, \ldots, G$ and $-100 \leq B^k \leq 100$. Next, to compute aggregate cell-type values from $\{B^k\}_{k=1}^{G}$, we 0 pad $B^k \in \mathbb{R}^{N_k \times N_k}$ to $\hat{B}_k \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}$ where $\hat{B}_{ij}^k = B_{ij}^k$ if both TF_i and TF_j are N_k^* final TFs (i.e. $c_i^*, c_j^* \neq 0$) selected for TG_k and $\hat{B}_{ij}^k = 0$ otherwise. Then, we compute the mean of these TG-specific undirected TF-TF coordination matrices $\{\hat{B}^k\}_{k=1}^{G}$ as $\mathcal{B}^{(0)} = \frac{\sum_{k=1}^{G} \hat{B}^k}{G}$ where $\mathcal{B}^{(0)} \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}$. We note there are $\ell_{\mathcal{N}} = 0.5\mathcal{N}(\mathcal{N} - 1)$ unique TF-TF links (i.e. $TF_i - TF_j$ link is the same as $TF_j - TF_i$ and hence counted once). We flatten $\mathcal{B}^{(0)}$ to a vector $\mathcal{B} \in \mathbb{R}^{\ell_{\mathcal{N}}}$ of corresponding values for the corresponding $\ell_{\mathcal{N}}$ unique TF-TF links. For each $TF_i - TF_j$ link, we convert $|\mathcal{B}_{ij}|$ to a percentile $0 \leq P_{ij} \leq 100$, where higher values indicate greater relative coordination among $\ell_{\mathcal{N}}$ links and $P \in \mathbb{R}^{\ell_{\mathcal{N}}}$; this enables a direct comparison of coordination strengths by normalizing $|\mathcal{B}|$'s distribution to a uniform 0 to 100 scale. We sort $|\mathcal{B}|$ in ascending order (small to large) and determine relative position, R_{ij}, of $|\mathcal{B}_{ij}|$ so $P_{ij} = 100\left(\frac{R_{ij}-1}{\ell_{\mathcal{N}}}\right)\%$ and higher P_{ij} reflects proportionally larger $|\mathcal{B}_{ij}|$. A pair of TFs may exhibit cooperative behavior to regulate some TGs and antagonistic behavior to regulate others. Our goal is to amalgamate learned coordination information across multiple TGs to reach a conclusion about the net nature of cell-type coordination (i.e. signed percentiles), in a manner akin to information content measures. We note that $\mathrm{sign}(\mathcal{B})$ is 1 if $\mathcal{B}_{ij} > 0$, -1 if $\mathcal{B}_{ij} < 0$, 0 otherwise. This informs the overall net cell-type coordination behavior between TF_i and TF_j.

The final cell-type TF-TF coordination network $\bar{B}$ is defined by: $\bar{B} = \mathrm{sign}(\mathcal{B}) \cdot P$, where $\bar{B}, \mathcal{B}, P \in \mathbb{R}^{\ell_{\mathcal{N}}}$ and $-100 \leq \bar{B} \leq 100$, is the same range as TG-specific coordination scores. Each non-zero score in $\bar{B}$ is unique. $|\overline{B_{ij}}|$ denotes relative strength of $TF_i - TF_j$ coordination and cell-type co-regulation (i.e. normalized interaction significance); the sign encapsulates their net observed behavior across G TGs: $-100 \leq \overline{B_{ij}} < 0$: both TFs are antagonistic overall, $0 < \overline{B_{ij}} \leq 100$: both TFs cooperate overall. TF-TG

regulatory networks provide more insights. If $\overline{B_{ij}} \approx 0$ and both TFs co-regulate TGs, we cannot deduce any meaningful information regarding the nature of their cell-type-specific interaction.

Thus, NetREm predicts how TFs coordinate with each other in networks to co-regulate TGs. For each of the G TGs in the cell-type, it outputs a TG-specific TF-TF coordination network: $-100 \leq B \leq 100$, where B_{ij} is the $TF_i - TF_j$ coordination to co-regulate TG. Here, $\bar{B}$ is a function f of $\{B^k\}_{k=1}^{G}$ across G TGs in the cell-type and is an overall cell-type-specific TF-TF coordination network: $-100 \leq \bar{B} \leq 100$, where $\bar{B}_{ij}$ is the net $TF_i - TF_j$ coordination behavior across all G TGs in the cell-type. The G TG-specific TF-TF coordination networks $\{B^k\}_{k=1}^{G}$ and the overall cell-type TF-TF coordination network $\bar{B}$ contain scores between -100 and 100, where negative scores suggest potential antagonistic co-regulatory relations and positive scores assign potential cooperative relations among the TFs, via direct and/or indirect mechanisms.

3.2.2 Real-world Datasets and Pre-processing

§B.1 (SaniyaKhullar 2024) details the parameters and evaluation for our applications. We applied NetREm for each TG in the cell-type for 7 main applications spanning 2 organisms (humans: 1, 2, 5-7; mouse: 3, 4). We ran NetREm without prior GRNs for these 5 applications: **1.** Simulated data for human embryonic stem cells (hESCs). **2.** Human Hematopoietic Stem cells (HSCs), which are self-renewing and long-lived cells in the bone marrow that are essential to produce blood cells (e.g. red blood cells, white blood cells, platelets). **3.** Mouse Embryonic Stem cells (mESCs), which are derived from the inner cell mass of the early embryo and are pluripotent since they can self-renew, develop, specialize, differentiate, and mature into any cell type in the body. **4.** Mouse Dendritic cells (mDCs), which are immune cells that capture and present antigens to other immune cells. **5.** Human Peripheral Blood Mononuclear cells (PBMCs), which contain 9 sub cell-types (i.e. clusters C) that are part of the immune system: Naïve CD4 T (C 0), CD14 Mono (C 1), Memory CD4 T (C 2), B cells (C 3), CD8 T (C 4), FCGR3A Mono (C 5), Natural Killer cells (C 6), Dendritic Cells (C 7), and Platelet cells (C 8). In these 5 applications, we fixed $N = \mathcal{N}$ candidate TFs for all TGs. If the TG is also a TF, we removed it from the set for the TG, so $N = \mathcal{N} - 1$. In

applications 1-4, we used gold standards to hone our TGs and TFs so TF-TG regulatory links are comparable with ground truth (McCalla et al. 2023); we trained models for TGs and for $\mathcal{N}$ TFs found in both gene expression and ground truth data. For PBMCs, we trained NetREm on 9 cell types with 1,029 TFs (Lambert et al. 2018).

We ran NetREm on 2 applications in humans using prior GRNs from multi-omics to define a custom set of N highly probable, candidate TFs for each TG (N differs across TGs): **6.** Myelinating (mSCs) and non-myelinating (nmSCs) human Schwann cells (SCs) in the peripheral nervous system (PNS). **7.** Alzheimer's disease (AD) and Control stages in humans for 8 cell-types in the central nervous system (CNS). 4 Glial cells: Astrocytes (Astro), Oligodendrocytes (Oligo), Oligodendrocyte Progenitor cells (OPCs or Oligodendrocyte Precursor cells); Microglia (Mic); 2 Neuronal cells: GABA-ergic Inhibitory neurons (InNs), Glutamatergic Excitatory neurons (ExNs); 2 Vascular and Blood-Brain Barrier (BBB) cells: Pericytes, Endothelial BBB (Endo. BBB) cells.

Single-cell gene expression data

Please note that **§B.1** (SaniyaKhullar 2024) provides preprocessing details for our 7 main applications and additional datasets, respectively. **1:** We randomly selected 1,250 TGs and corresponding TFs from weighted and signed (+: activates; -: represses) ground truth GRN atlas from TF induction analysis (Sharov et al. 2022). This results in $\mathcal{N}$ = 207 TFs and 5,050 GRN links we input to SERGIO to simulate realistic single-cell data for 100 and 1,000 cells and 1,442 genes. We varied the noise parameter settings (30, 60, 90)%, retrieving 3 different synthetic expression datasets. **2:** We used (Buenrostro et al. 2018). **3:** We used (Tran et. al 2019) that reprograms mouse embryonic fibroblasts to embryonic-like induced pluripotent stem cells. **4:** We used normalized data (McCalla et al. 2023) from (Shalek et. al 2014) for >1.7k primary bone marrow DCs. **5:** We used public healthy donor data on 2.7k PBMCs based on (Satija et. al. 2024). **6:** We used (Avraham et al. 2022) for mSCs and nmSCs in DRG L4,5 regions from 5 donors. **7:** We used processed (Gupta et al. 2022) for 24 AD and 24 healthy humans for 8 cell types based on 80,660 droplet-based single-nucleus prefrontal cortex transcriptomes (Mathys et. al 2019).

Prior GRN reference information on TG regulation

In applications 6-7, we employed multi-omics and epigenetic data from open scATAC-seq chromatin regions mapped to TGs peak-TG links) to identify potential interacting REs for TGs. We mapped sequence-specific TF motifs to REs, using Position Weight Matrix (PWM) databases to predict TFBSs, forming a motif-based GRN: direct TF-RE-TG candidate links. We pruned this motif-based TF list based on relative TF expression (proxy for protein abundance) and motif matching scores. To capture overlooked TFs, we augmented our pruned TF list by adding TFs with known PPIs and predictions of TF-TF colocalization, complexes, and/or Molecular Function similarity (Wu et al. 2021). This addresses limitations of GRN-inference tools that rely solely on accessible motif matches and may miss causal TFs for TGs (Zhang et al. 2023); as ~10% of the ~1.6k human TFs lack motif data and are traditionally excluded, using PPIs during GRN inference is recommended to incorporate these missing TFs (Badia-i-Mompel et al. 2023). Augmented TFs may bind to TFBSs directly (weak signals) or indirectly (associate with DNA-binding TFs) (Gordân et al. 2009; Sloan et al. 2016). Our final prior GRNs comprise these initial TF-RE-TG links. We input the N biologically-promising TG-specific candidate TFs to NetREm for the given TG.

These subjective steps are detailed in **§B.1** (SaniyaKhullar 2024): step-by-step. We showed how we integrate these data sources (along with others) to help determine the list of N candidate TFs for each target gene (TG), which will be part of our input gene expression data X when running NetREm for each TG in the given cell-type. We note that data provided by recent studies enables the construction of prior gene regulatory networks for various cell-types in the human body (e.g. (Zhang et al. 2021) that uncovers regions of open chromatin in 222 distinct human cell-types; scQTLbase (Ding et al. 2023) currently integrates single-cell eQTL data for 57 different cell types and 95 cell states). Incorporating this prior information (e.g. TF binding predictions, chromatin accessibility) from multi-omics data can lead to more biologically meaningful and potentially truer cell-type TF-TG regulatory networks (Badia-i-Mompel et al. 2023).

Organism-Specific Protein-Protein Interaction (PPI) Networks (PPINs)

We employed public STRINGdb (Szklarczyk et al. 2023) to construct human and mouse PPINs (details: **§B.1** (SaniyaKhullar 2024); we add more resources for humans) for proteins that partner through direct (physical binding, complex coexistence) and/or indirect (e.g., metabolic/signaling paths) PPIs (Oughtred et al. 2021) based on the full networks of scored links between proteins in the protein network data. We scaled average combined scores (across many evidence types) to assign weights: $0.01 < w \leq 1$. Any self-loops are ignored. For each of the G TGs in a context, we filtered this partly-connected PPIN $\mathbb{W}$ to only keep its N TFs to yield $W^{(0)}$. Nonetheless, some of the N TFs may not be in $W^{(0)}$ and some edges may be missing for the existing TFs for existing TFs in this potentially partly-connected $W^{(0)}$. Thus, when we run NetREm, we assigned an artificial weight $\eta = 0.01$ to missing edges so $0.01 \leq w \leq 1$ for $i \neq j$ for all N TFs. This ensures numerical stability and propel discovery of novel TF-TF links; this yields our TG-specific final comprehensive fully-connected TF-TF PPIN (input W).

§ 3.3 Results

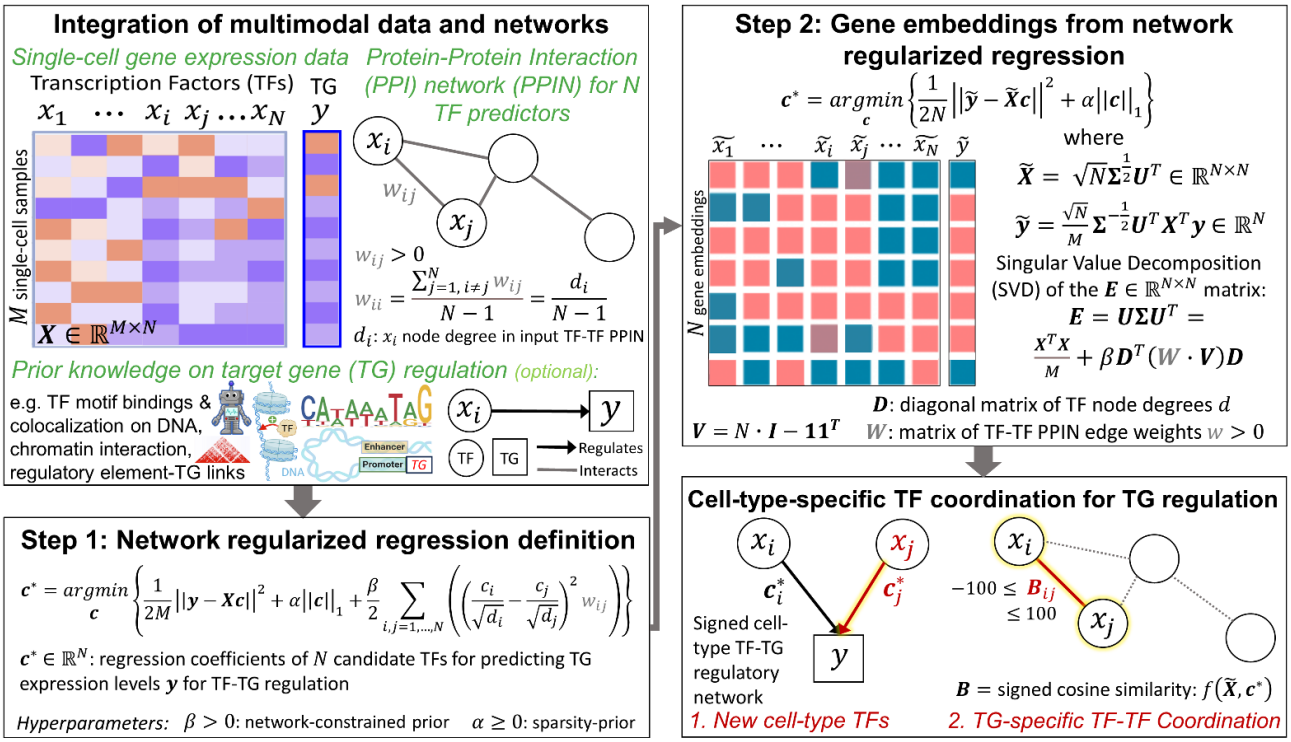
3.3.1 Overview of Network Regression Embeddings (NetREm)

We developed NetREm as a robust, multi-omics, computational approach to build and integrate networks of TF-to-TG regulation and TF-TF coordination in a cell-type-specific manner (**Figure 3.1**). Network regularization-based approaches like NetREm are useful in applications where predictors form subnetworks to influence the outcome (Li and Liu 2022). NetREm uses single-cell or bulk gene expression data, a comprehensive PPIN (e.g. STRINGdb) that globally captures verified direct/indirect functional associations among TFs (Peng et al. 2017). Providing a prior, candidate GRN (e.g. initial TF-RE-TG links) is optional but recommended as it improves biological relevance of outputs with its initial feature selection; TFs in this prior GRN are considered candidates for the TG and NetREm identifies the subset of these links that are predictive. NetREm integrates this multi-modal data, applying a 2-step optimization process for each TG: **1.** It formulates a network-regularized regression problem, using the input PPIN as a priori information, to sift through N candidate TFs, to find those most likely to co-regulate TG. The input PPIN is processed to become a fully-connected network comprising not only known but also artificially

added links (with minimum edge weight w) for the N candidate TF proteins for the TG. **2.** To solve this regression problem, it then employs Singular Value Decomposition (SVD) to create network regression embeddings, which are used in Lasso regression to predict TG expression. In the process, NetREm learns not only key TFs (coefficients c^*, a complementary GRN) but also a TG-specific TF-TF coordination network $-100 \leq B \leq 100$, predicting cooperative (> 0) or antagonistic (< 0) TF-TF relationships for regulating TG. It also annotates known PPIs, uncovering confident cell-type-specific TF-TF PPI subnetworks involved in TG regulation (Szklarczyk et al. 2023). By integrating genomic and foundational PPIs, NetREm uses transfer learning to discover unanticipated, biologically significant TF-TG links and novel functional TF-TF relationships for future investigation (Zhang et al. 2013). Aggregating outputs across all TGs, yields cell-type-specific directed TF-TG regulatory and undirected overall TF-TF coordination ($-100 \leq \bar{B} \leq 100$) networks. These outputs support downstream analyses, like linking non-coding SNPs to potential regulatory roles via expression trait loci (eQTL) SNP-TG (i.e. eSNP-eTG) associations in our networks.

We demonstrated NetREm's versatility across diverse real-world scenarios by benchmarking our TF-TG networks against established gold standard GRNs in human (hESCs, HSCs) and mouse (mESCs, mDCs) models. We used diverse techniques to assess $\bar{B}$ across these settings, as well as additional human contexts: 9 PBMCs, pooled stem cells from GTEx (Eraslan et. al 2022), 4 CNS cell types (Lake et al. 2018). Further, we highlighted NetREm's use of prior input GRNs, derived from multi-omics data, for specific contexts (cell types, diseases) in 2 human applications: 1. mSCs vs. nmSCs; 2. AD vs. controls in 8 neuron/glia cell types. The resulting TF-TG networks, integrated with TF-RE-TG annotations (**Methods: §3.2**), yield enhanced, context-specific TF-RE-TG regulatory and coordination $\bar{B}$ networks.

Figure 3.1 – Overview of Network Regression Embeddings (NetREm) a multi-step, multi-omics computational framework to construct comprehensive cell-type-specific directed TF-TG regulatory network and undirected TF-TF coordination $\bar{B}$ networks.



We apply this pipeline for each TG in a cell-type of G TGs. NetREm integrates multimodal data: gene expression ($X \in \mathbb{R}^{M \times N}$, $y \in \mathbb{R}^{M}$) for M cells (samples), direct/indirect TF-TF PPIs with weights W, optional prior GRN information (e.g. TF-RE-TG links). This prior GRN helps select only relevant TF predictors for TG (from multiomics data) to reduce dimensionality of X: M (rows), N TFs (columns). X and y are standardized (mean μ: 0, standard deviation σ:1) across cells for TFs and TG. Goal: identify TFs, out of N candidates, whose expressions best predict TG expression y. (**Methods** provides details)

- **Step 1** involves setting up a PPI network-regularized regression problem (preprocessing it with artificially added minimum edges to ensure it is fully-connected for N candidate TFs) to identify optimal TFs (out of the N) for TG, guided by: network prior hyperparameter β, sparsity prior hyperparameter α.
- **Step 2** solves this, transforming (X, y) and PPIs to latent space gene expression embeddings ($\tilde{X} \in \mathbb{R}^{N \times N}$, $\tilde{y} \in \mathbb{R}^{N}$) by SVD on an $E = \frac{X^T X}{M} + \beta A$ matrix ($A = D^T(W \odot V)D$). E combines expression relations and PPIN information. $(\tilde{X}, \tilde{y})$ undergo Lasso regression (via model-type: Lasso (α is given) or LassoCV (α chosen by cross-validation (CV); default: no intercept) to predict optimal coefficients $c^* \in \mathbb{R}^{N}$ for TG.

NetREm outputs 2 networks capturing different aspects of TG regulation that can be integrated. #1: links optimal TFs to TG by c^* (> 0: activator, < 0: repressor) in a TF-TG regulatory network. This complementary GRN likely uncovers novel cell-type TFs like x_j and reflects underlying biology of TF-TF coordination. If we input a prior GRN, we may use it to annotate our TF-TG links with biological metadata (e.g. TF-RE-TG network). #2: TG-specific TF-TF coordination network B predicts indirect/direct relationships among TFs to regulate TG. $TF_i - TF_j$ coordination B_{ij} shows cooperative (> 0 if both TFs are co-activators or co-repressors) or antagonistic: (< 0 if 1 is repressor, other is activator) co-regulation of TG. Here B: c^*-aware cosine similarity scores, a function f of $\tilde{X}$ and c^*. NetREm thus identifies novel coordination among cell-type TFs for co-regulating TG. Results are stitched together across runs for all G TGs to obtain final cell-type outputs, including the overall cell-type TF-TF coordination network $\bar{B}$.

3.3.2 Simulation study

We tested NetREm on simulated single-cell gene expression data for 10,000 cells as a proof-of-concept, where: X (matrix of expression levels for N = 5 candidate TF predictors) and y (TG expression vector) were drawn from a normal distribution with dropouts to achieve ~40% sparsity, mimicking single-cell data (**Figure B.1A** (SaniyaKhullar 2024)). Training data (70%; M: 7,000 cells) standardizes (X_{train}, y_{train}, X_{test}, y_{test}) so that each variable has ($\mu = 0$, $\sigma = 1$). TFs 1 to 5 have expression with Pearson $r(TF, TG) \approx$ [0.9, 0.5, 0.4, -0.3, -0.8] with y, respectively, in training and testing data (**Figure 3.2A, Figure B.1B-B.1C** (SaniyaKhullar 2024)). Our PPIN of known direct and/or indirect PPIs among TFs has strong w for $TF_1 - TF_2$ (0.8), $TF_4 - TF_5$ (0.95). We set w to $\eta = 0.01$ for missing PPIs, making our PPIN fully-connected. NetREm(β = 1, α = 0.1) predicts y based on X, subject to the PPIN constraint. It outputs 2 networks: TF-TG regulation (c^*) relates expression levels of TFs to TG expression (**Figure 3.2B**) and TG-specific TF-TF coordination (B scores) (**Figure 3.2C**) of behavior among TFs to regulate TG.

We compared NetREm to 4 default Scikit-Learn (Pedregosa et al. 2011) benchmark regression models (BRMs) fit with no intercept: Linear Regression and 3 regularization ones (ElasticNetCV, LassoCV, RidgeCV). TFs with lower ranks have higher $|c^*|$ and are more important. Absolute values of c^* for TF_2 and TF_4 are significantly higher in NetREm (**Figure B.1D, Table B.1:** p < 2e-16 (SaniyaKhullar 2024)), highlighting NetREm's grouped variable selection property that prioritizes them due to their strong corresponding PPIs with TF_1 and TF_5 that both strongly correlate with y (Li and Li 2008) (**Figure B.1E** (SaniyaKhullar 2024)). Unlike BRMs that favor TF_3 over TF_4 (since TF_3 has greater $|r|$ with y), NetREm prioritizes TF_4 over TF_3 since TF_4 strongly interacts with TF_5, while TF_3 has weak PPIs with other TFs. Sensitivity analysis (**Figure 3.2D-E**) for fixed α and β confirms $|c^*|$ for TF_2 and TF_4 increase as their respective PPIs with TF_1 and TF_5 strengthen. NetREm's test MSE increases from 0.15 to 0.22 as β increases from 0.01 to 1 (**Figure B.1F** (SaniyaKhullar 2024)). Test MSEs for BRMs ≈0.14, which we can also achieve with NetREm(β = 0.01, LassoCV α); nonetheless, studies emphasize

obtaining more interpretable and context-driven features by incorporating network information into models, even if accuracy is sacrificed (Li and Liu 2022). **Figure B.1G** (SaniyaKhullar 2024) illustrates changes in B as β increases from 0.01 to 2.

One of NetREm's salient features is its ability to explicitly model and integrate PPIN structures and relationships among TFs. By doing so, it identifies key TFs for predicting y, which may not exhibit strong r with y individually. NetREm also generates a coordination network among TFs that co-regulate TG, a capability beyond the scope of BRMs. Instead, BRMs, primarily focus on prediction accuracy by selecting c^* to best represent each TF's impact on y, typically treating TFs as "lone wolves", neglecting TF-TF interdependencies vital to GRNs. Our B network shows TF_1 and TF_2 cooperate with TF_3 weakly to increase y, TF_4 and TF_5 collaborate to decrease it; TF_1-TF_2 and TF_4-TF_5 links are highly-confident. NetREm predicts novel direct/indirect links like TF_1-TF_5 activator-repressor antagonism B = -97.9. TF_3 has weaker relations with TF_1 than TF_2 does (41.7 vs. 71.5) and the smallest $|c^*|$ = 3.4e-2.

We detailed more simulations in **Figure B.2A-F**, **Figure B.3A-C**, **Figure B.4A-H**, and **Figure B.5A-I** and **Tables B.2-B.3** in **Chapter §B** (SaniyaKhullar 2024)). We showed that results are stable and consistent for various expression sparsity levels and for $M \ll N$ cases as well, excess β over-constrains NetREm causing its predictions to suffer (bias-variance trade-offs), NetREm has more robust c^* estimates (less variable) and more accurate TF assignments, and it can capture complex TF-TF PPIs. Our toy data with a few TFs and 1 TG intuitively explains NetREm and its advantages.

Overall, NetREm's network constraint regularization enhances model robustness as it is less sensitive to noise and artifacts in the data. This is especially beneficial in high-dimensional datasets (e.g. gene expression data, where there are more features (genes) than samples) where the risk of overfitting and multicollinearity are prevalent; this may result in correlations among variables that are simply by chance and not inherently meaningful (Campos et al. 2019; Hoefsloot et al. 2008). Further, single-cell gene expression data is susceptible to high dropout and sparsity (e.g. ≈70% of entries can be close to 0), such that the true distribution of gene expression of genes may not be captured and results may be unreliable

(Nguyen et al. 2020). Sometimes, this data can be noisy due to the inherent complexity of gene regulation in eukaryotes (e.g. humans), which may obscure TF-TF correlations, making it difficult to identify functionally coordinated TFs (Nie et al. 2011). Still, TFs typically co-regulating several common TGs (i.e. have strong synergistic activity and behavior) may potentially exhibit high correlations with each other than with other TFs due to their similar and loosely coordinated gene expression profiles (Roy et. al 2020; Nie et al. 2011). This high intercorrelation among TFs may be problematic when using gene expression data to predict TF-TG regulatory links. For instance, many correlated TF predictors may truly be related to TG expression (e.g. TFs GATA1 and P300 are quite correlated and form a complex for TG regulation in K562 cell lines (Ahsendorf et al. 2017)); however, benchmark (and other machine learning) models tend to struggle with highly correlated features, and usually select independent TF predictors or drop a few of these true co-regulating TFs. This sadly compromises the integrity of the learned TF-TG regulatory network (Roy et. al 2020) as some correlated TFs may be causal for TG regulation (Nicodemus and Malley 2009) and many TFs typically coordinate to regulate TG expression (Ibarra et. al 2020). Coordinating TFs are likely to share common PPIs, including potentially strong, direct and/or indirect physical interactions with each other as partners to regulate shared TGs (Perna et al. 2020).

NetREm's network-constrained regularization optimization problem incorporates the structure of these strong, known, indirect and/or direct physical PPIs among TFs (given by respective PPI weight), to encourage coefficients among strongly connected TFs in PPI network to be more similar. Integrating this PPI may account for some of the noise in gene expression data, which may be attributed to TF coordination (Parab et al. 2022). Thus, by leveraging network pre-knowledge, NetREm accurately discerns and assigns each TF's influences, yielding dependable, detailed insights into individual roles, undeterred by intercorrelations and instead inspired by a promising ground truth framework. NetREm offers superior generalizability and consistency, potentially capturing intricate underlying structures and interactions in the data that might be oversimplified by benchmark models. This balanced integration of predictive accuracy and structural interpretability positions NetREm as a computationally efficient and methodologically sound choice. In summary, NetREm integrates TF-TG regulatory links and TF-TF

coordination scores, holistically, to understand better the TF behaviors involved in regulating TG and deepen our grasp of TF regulatory dynamics. Next, we benchmarked NetREm in real-world settings: many TFs and TGs. We applied it step-by-step for each TG in a cell-type to learn B (i.e. TF-TF coordination network for each individual TG in the cell-type), cell-type TF-TG regulation, and overall cell-type TF-TF coordination network $\bar{B}$. Here, $\bar{B}$ is a function f of $\{B^k\}_{k=1}^{G}$ across G TGs in the cell-type.

Figure 3.2 – Simulation study of 5 TFs and 1 target gene (TG)

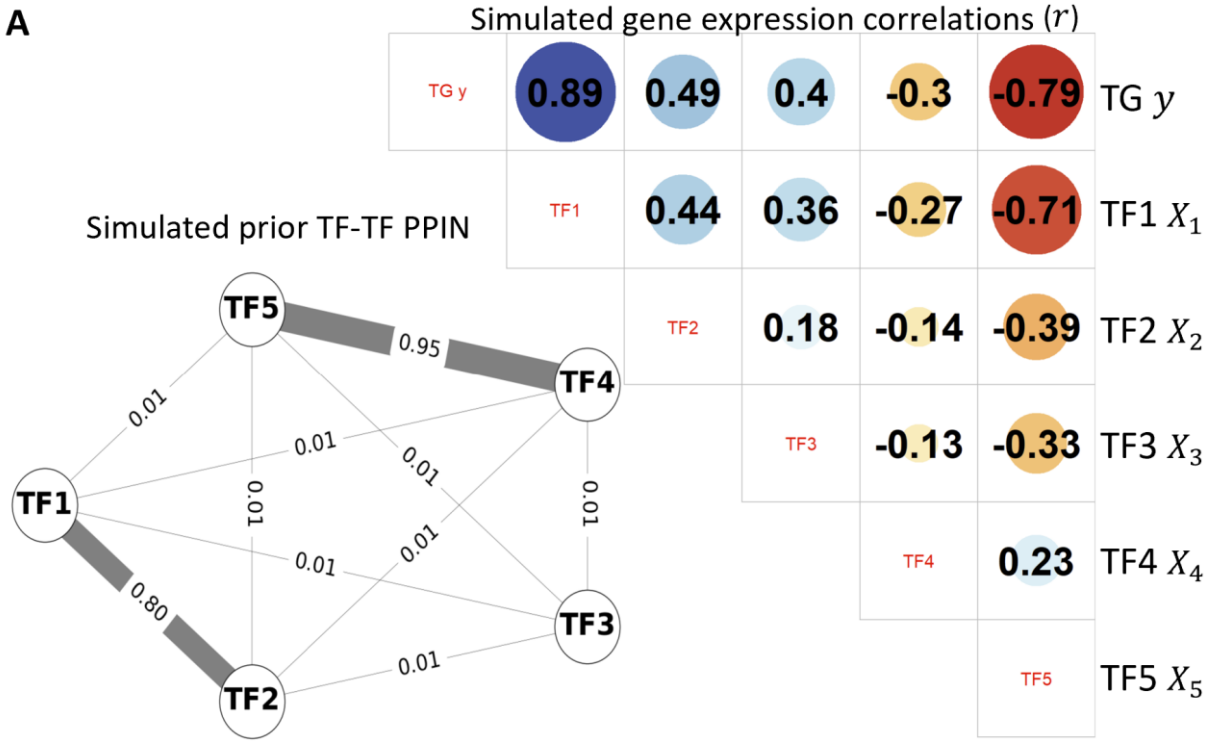


Figure 3.2A) - The bottom left shows the prior biological, undirected TF-TF PPI network with default edge weights (0.01), with stronger experimentally-verified connections for TF_1- TF_2 (0.8) and TF_4- TF_5 (0.95). The top right presents a Pearson correlation (r) matrix among the TFs and TG in the training gene expression data where: $r(TF, TG)$ = [0.9, 0.5, 0.4, -0.3, -0.8]. Dot sizes represent magnitude, and colors indicate positive (blue) or negative (red) correlations.

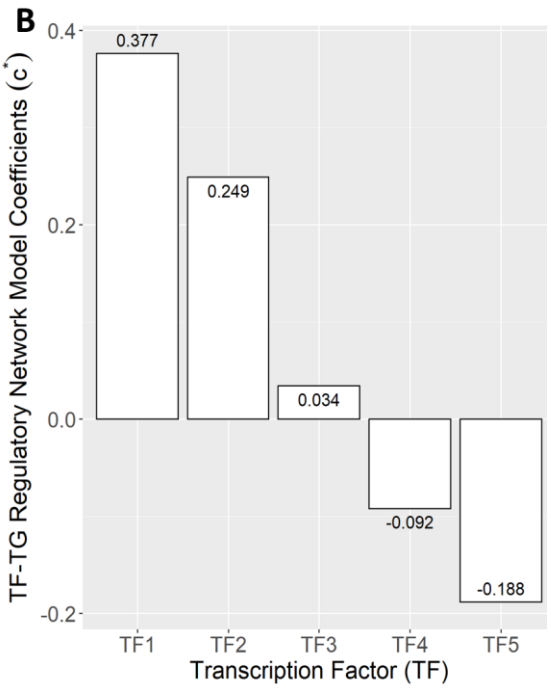


Figure 3.2B) - Coefficients c^* for TFs in TF-TG regulatory network, based on NetREm(β =1, α = 0.1, no y-intercept). Potential activators: TF_1 to TF_3. Repressors: TF_4 and TF_5.

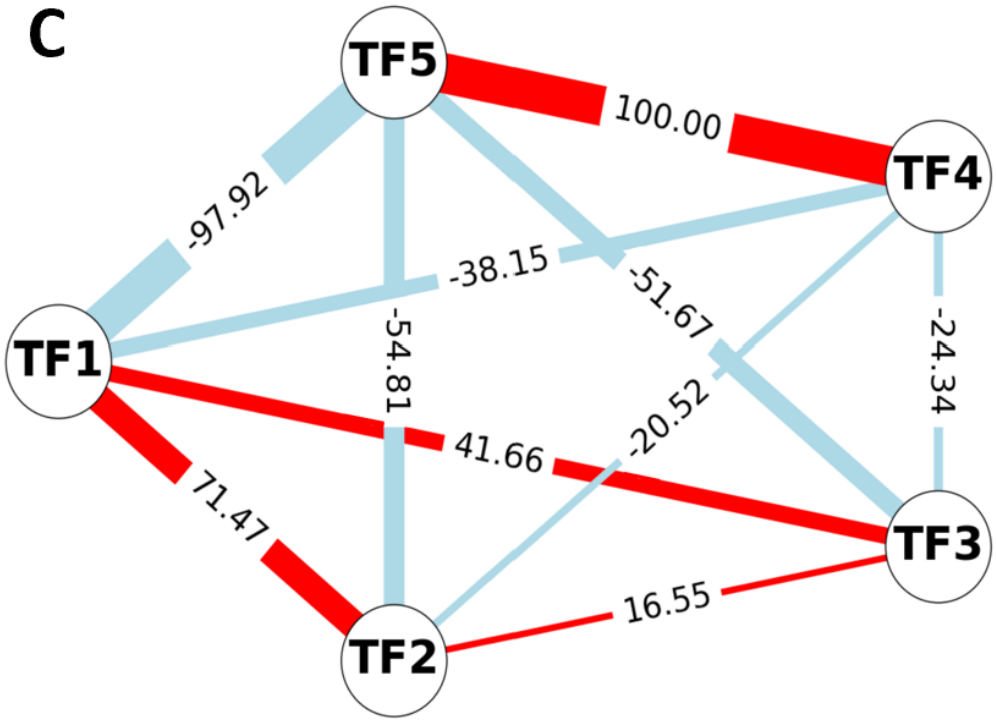


Figure 3.2C) - TG-specific TF-TF coordination network (scores: B) with red (antagonistic: -; activator-repressor links) and blue (cooperative: +; links between co-activators or co-repressors). Functionally-valid coordination: TF_1- TF_2 and TF_4- TF_5; others are novel.

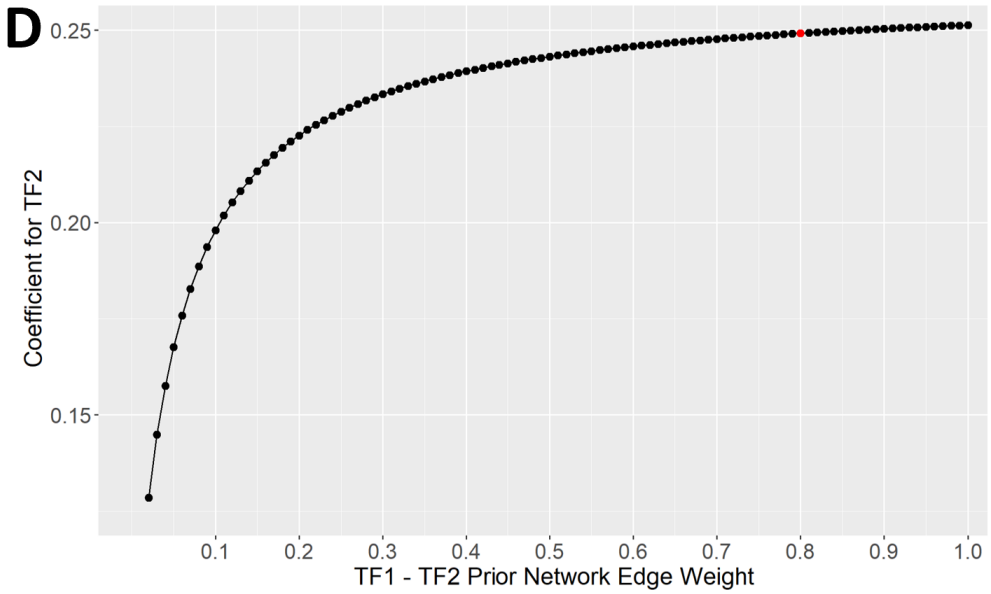


Figure 3.2D) - Effects of varying TF_1-TF_2 w in original PPIN, from 0.01 to 1 in 0.01 increments, holding all else fixed (e.g. same expression data and NetREm(β =1, α = 0.1, no y-intercept)) Respective TF_2 c^* coefficient increases monotonically in arc shape from 0.106 to 0.251. The red dot signifies the selected edge weight used for the simulated TF-TF PPI network (i.e. the w).

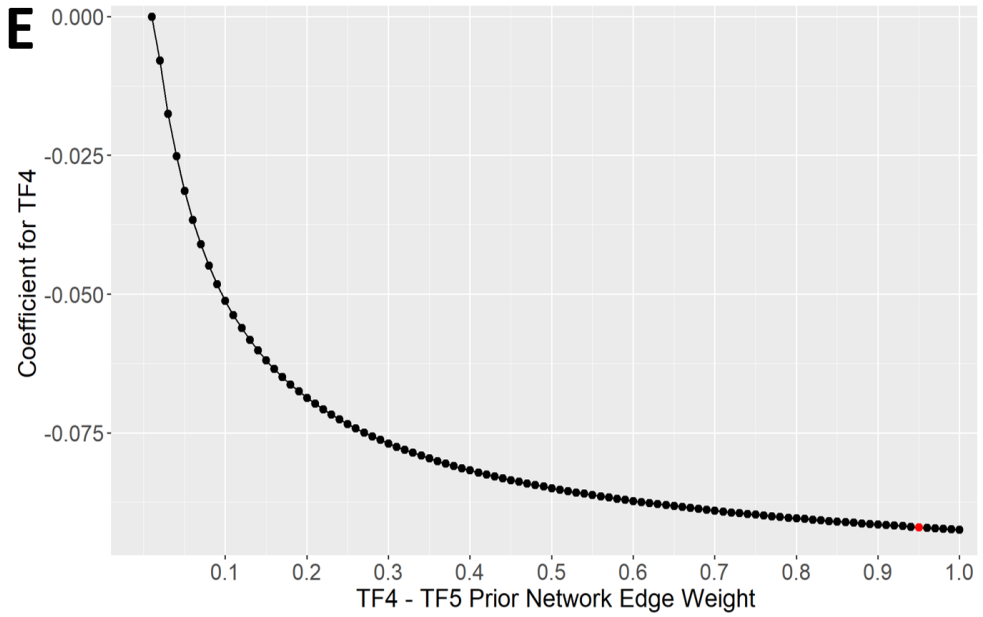


Figure 3.2E) - A similar sensitivity analysis for TF_4 and TF_5 shows that TF_4's $|c^*|$ increases, becoming more negative from 0 to -9.2e-2, as TF_4-TF_5 edge weight is perturbed.

3.3.3 Benchmarking NetREm with No Prior GRN Information

Now we assess how effectively NetREm predicts cell-type TF-TG and $\bar{B}$ networks in real-world scenarios lacking prior GRN information (**Methods: §3.2**). All TGs have the same N candidate TFs; for TGs that

are TFs, N is reduced by 1. We anticipate that incorporating prior GRNs derived from multiomics data may enhance our performance by providing a tailored, context-specific list of promising TFs for each TG.

3.3.3.1 Evaluating TF-TG regulatory links

To evaluate our networks in terms of predicting regulatory TF-TG links as well as accurate TF roles (based on signed c^*: + for activators, - for repressors), we compare NetREm with BRMs for predicting signed-TF-TG links. We use SERGIO (Dibaeinia and Sinha 2020) to input a signed hESC GRN and generate 6 realistic datasets (1,250 TGs, N = 207), varying M and noise %. Overall, NetREm has the highest precision, indicating its superior reliability in correctly identifying true TF-TG links, assigning activator/repressor roles, minimizing FPs (**Figure B.6A-F** (SaniyaKhullar 2024)).

We benchmark our inferred TF-TG links using single-cell expression data for human HSCs and mice (mESCs, mDCs) (**Figure B.7A-I** and **Figure B.8**, **Tables B.4-B.6** (SaniyaKhullar 2024)) against gold standard GRNs (McCalla et al. 2023; Zhang et al. 2023). Due to its use of biological information (Shojaie and Michailidis 2009), NetREm demonstrates higher sensitivity in identifying relevant biomarkers, though it shows lower specificity compared to Lasso and ElasticNet (Li and Li 2008). By adjusting β and α, we fine-tune NetREm's behavior; increasing α might decrease sensitivity but enhance specificity. Our findings are consistent across updates in input PPINs, comparing well between mouse STRINGdb versions 11 and 12 in mESCs.

NetREm features a unique grouped variable selection mechanism that leverages PPIN structures to prioritize tightly-linked TF groups involved in known PPIs. This optimizes the selection of cell-type TFs that co-regulate TGs, enhancing NetREm's ability to identify complex TF-TF relations. We highlight this in HSCs. ElasticNet and Lasso fail to identify critical TF-TG links, for instance, missing regulation of *ATF2*, a key HSC TF (Ju et al. 2023). Conversely, Linear Regression and Ridge predict that *ATF2* is regulated by all N = 177 TFs, showing their alarming potential for FPs. However, NetREm(β = 10) identifies 8 TFs for *ATF2*, with 7 confirmed by gold standards. This pattern extends to other TGs like *BRD2, RNF167, DUSP2*; NetREm identifies groups of both validated and novel TFs involved in

biologically relevant PPIs for these TGs and more. For instance, 1 of 10 novel TFs for *RNF167* is TFAP4, which forms PPIs for adipogenesis and cell population proliferation regulation alongside 3 of 17 verified final TFs. NetREm effectively uses prior TF-TF PPIN knowledge to identify genuine TF-TF coordination for TG regulation, navigating through data noise and avoiding false correlations (r). It integrates r and PPIs among TFs, to achieve superior generalizability and consistency in identifying intricate TF-TF relationships that BRMs may struggle with.

Our TF-TG networks complement state-of-the-art cell-type GRN tools. To show this, we utilize input (normalized gene expression data), validation (ground truth GRN for evaluating accuracy of the inferred GRNs), and output (predicted state-of-the-art (SOTA) cell-type GRNs) data provided by a previous study (McCalla et al. 2023) that benchmarked the performance of these SOTA GRN inference tools in mouse dendritic cells (mDCs): Inferelator (Greenfield et. al 2013), knnDREMI (van Dijk et. al 2018), LEAP (Specht and Li 2017) mean, MERLIN (Roy et al. 2013), Pearson mean (baseline network: undirected fully connected network, where edges are weighted by correlation between each pair of genes over all cells), PIDC (Chan et al. 2017) mean, SCENIC (Aibar et. al 2017), SCODE (Matsumoto et al. 2017) mean, Scribe (Qiu et al. 2020), SILGGM (Zhang et al. 2018) mean. These SOTA GRN inference models utilize the following methods for network inference from scRNA-seq data: graphical models and dependency networks (Inferelator, MERLIN, SCENIC, SILGGM), information theoretic (kNN-DREMI, PIDC, Scribe), ordinary differential equations (SCODE), correlation (LEAP, Pearson) (McCalla et al. 2023). We compare the performance of NetREm's TF-TG regulatory network in mDCs to those of these SOTA GRN inference tools. NetREm has comparable performance with these SOTA GRNs across various metrics (sensitivity, specificity, F1 Score, balanced accuracy, overall accuracy). No method outperforms another for predicting TF-TG links. NetREm, however, infers TG-specific B and cell-type-specific TF-TF coordination networks $\bar{B}$, capabilities that other tools lack.

3.3.3.2 Evaluating TF-TF Coordination

We evaluated the performance of our $\bar{B}$ in mESCs, mDCs, and PBMCs using V11 PPIN as input (proxy for outdated information), categorizing TF-TF pairs into 4 groups based on their status in V11 and in updated V12: absent in both, present in both (TPs), removed in V12 (FPs), discovered in V12 (FNs). Top TF-TF links have higher $|\bar{B}|$. Welch 1-sided tests (p-adj < 0.05) compare $|\bar{B}|$ across groups (**Tables: B.7-B.12; Figures: B.9A-D**, **B.10A-B**, **B.11A-C** (SaniyaKhullar 2024)). For instance, in mESCs, NetREm often reflects known PPIs, flags FPs to remove, and uncovers biological truths, nominating promising candidate PPIs for follow-ups. Overall, NetREm prioritizes known TPs and can potentially flag future TF-TF PPIs (FNs) that are currently unknown.

Further, we benchmarked NetREm against the RTNduals (Chagas et al. 2019) tool. In 12 of 13 human contexts (**Figure B.12A-C** (SaniyaKhullar 2024)), our top k links using V11 input PPIN outperform RTNduals, with a higher % that are verified PPIs in V12 and other sources. This underscores NetREm's efficacy in identifying TPs and leveraging historical PPINs to predict previously unknown PPIs; this is encouraging as only a small number of ~130-650k estimated human PPIs are currently known (Sevimoglu and Arga 2014; Venkatesan et al. 2009; Yu et al. 2020).

To validate NetREm's capability to prioritize biologically relevant links, we leveraged a Contextual PPI Database (CPPID) (Kotlyar et. al 2022) that annotates PPIs with >243 terms but lacks cell-type specificity (**Figure B.13** (SaniyaKhullar 2024)). By showing that our top links are indeed enriched for context-specific terms (e.g. nervous system (NS)-related in Microglia (Mic) and pooled SCs, immune-related in PBMCs), we suggested a potential extrapolation to cell-type specificity. This highlights NetREm's pioneering potential to discover cell-type TF-TF PPIs, addressing a crucial gap in existing global PPINs.

3.3.4 Gene regulatory links between transcription factors (TFs) and target genes (TGs) in myelinating and non-myelinating human Schwann cells (SCs)

With emerging new single-cell epigenomic data from many human tissues, we can model GRNs in novel contexts. We applied NetREm to analyze SCs, which are pivotal in maintaining, regenerating,

myelinating, supporting PNS neurons. SCs are derived from neural crest and exhibit tremendous flexibility not only in myelination but also in several other tissues; they also function as terminally differentiated cells that can reverse differentiation after nerve injury to aid nerve regeneration (Ma and Svaren 2018). Single-cell rodent NS studies reveal substantial diversity in SC differentiation status (Gerber et al. 2021; Yim et. al 2022). 2 main SC phenotypes (or sub-types) are: 1) mSCs associated with larger diameter axons (>1 micron); 2) nmSCs that wrap a bundle of smaller diameter axons (typically sensory axons) i.e. a Remak bundle. While mSCs envelop axons in myelin sheaths to enhance conduction speed, nmSCs support sensory axon function/interactions without forming myelin, contributing to overall nerve integrity.

Uncovering TF-TG regulatory mechanisms modulating cellular processes in mSCs and nmSCs may help us understand and treat debilitating nerve injuries, demyelinating disorders, and hereditary neuropathies. Mutations affecting SC function are the most prevalent cause of demyelinating genetic neuropathy Charcot-Marie-Tooth disease (CMT) (Tao et. al 2019) and some affect major transcriptional TFs of SC differentiation like EGR2 and SOX10 (Srinivasan et al. 2012; Fröb and Wegner 2021). Both TFs are co-expressed in myelinating SCs, colocalizing at several myelin REs (Jones et al. 2007; Poitelon et. al 2016). Mutations in EGR2 may disrupt cooperative SOX10 TFBS binding (LeBlanc et. al 2006) and its subsequent regulation of TGs. SC regulation also involves TEAD1 and other Hippo regulators to govern shared TGs and orchestrate PNS myelination (Srinivasan et al. 2012; Lopez-Anido et al. 2016). Since there are no TFs that are exclusively expressed by SCs, uncovering TF-TF coordination networks crucial for SC lineage maturation and differentiation into mSCs or nmSCs also holds significance (Ma and Svaren 2018). However, distinct transcriptional regulatory networks coordinating SC function and underlying states and cell fates require a variety of TFs beyond EGR2 and SOX10 (Hung et al. 2015).

In response, we applied NetREm to each TG in mSCs and nmSCs using single-cell data for human Dorsal Root Ganglion (DRG). We derive prior GRNs using multiomics data (details: **Methods: §3.2**). To do this, we annotate open chromatin regions in adult human SCs with known RE peak-to-TG links (Zhang et al. 2021) and use motif-based analysis to predict TFs that may associate with these REs. By

removing low-expressed TFs in corresponding training gene expression data for mSCs and nmSCs in the human DRG, we create tailored prior mSC and nmSC GRNs. For each TG in an SC-type, we input its N TG-specific selected candidate TFs (from respective prior GRN of initial TF-RE-TG links) to NetREm. In total, NetREm outputs 183,242 mSC and 277,541 nmSC total TF-TG links (**File B1** (SaniyaKhullar 2024)) comprising 221 TFs and 8,950 TGs in mSCs, 228 TFs and 5,207 TGs in nmSCs. Both share: 33,806 TF-TG links, 27,037 sign-TF-TG links, 3,841 TGs, 197 TFs. TF EGR2 is mSC-specific (Balakrishnan et al. 2021). We enhance our networks by overlaying them with prior GRN annotations, resulting in our finalized TF-RE-TG regulatory networks.

We examined NetREm's results for 8 core SC TFs that have validation (genome binding and loss-of-function (LOF)) data from rodent SCs: EGR2, NR2F2, RXRG, SOX10, SREBF1, STAT1, TEAD1, YY1. LOF TGs, whose expression varies upon TF knockdown, show their direct or indirect dependency on the TF (Nie et al. 2020). Valid direct TGs are a subset of LOF TGs with ChIP-seq evidence of nearby TF binding, suggesting direct regulation by the TF (Titsias et al. 2012); this dual confirmation enhances confidence in these direct TGs (Badia-i-Mompel et al. 2023). RXRG lacks LOF data. In **Figure 3.3A** we provide counts of NetREm-predicted: direct, all LOF, and novel candidate TGs. We also report on eTGs with strong Tibial Nerve eQTL support (THE GTEX CONSORTIUM 2020) based on mapped SNP-TF-RE-TG predictions: instances where an eSNP strongly alters TF binding to a RE, influencing TG expression (**Methods: §3.2**). For example, we predict 2,015 YY1 TGs in nmSCs (139 direct, 304 LOF only, 1,572 novel), where LOF and direct TGs are significant (hypergeometric $p < 0.05$); of these, 40 direct (28.8%), 75 LOF only (24.7%), 488 novel YY1 (31%) TGs are eTGs. In both SCs, our final TF-TG predictions for all LOF and direct TGs have higher sensitivity and F1 scores compared to LassoCV and ElasticNetCV and higher accuracy and specificity than GRNBoost2, Linear Regression, RidgeCV (**Figure B.14** (SaniyaKhullar 2024)).

We explore cases where NetREm accurately identifies a core SC TF as the top predictor (low rank) for its direct TG, even when training expression data shows a low r(TF, TG) (**Figure 3.3B**), i.e. low correlation between TF and TG gene expression levels. This is important as studies observe that in

eukaryotes (unlike in prokaryotes), Pearson correlations r and mutual information among TFs and known TGs are not much higher than those between TFs and non-TGs (Zaborowski and Walther 2020; Escorcia-Rodríguez et al. 2023). For example, SOX10 weakly correlates with *SIPA1L2* (Tao et. al 2019)(candidate modifier TG for CMT type 1A) but predicted to have rank 17. Despite its relatively weak r = -0.14 with *MBP*, a major constituent of myelin sheaths, STAT1 is its top 10 TF. Although *APP* exhibits a higher r with STAT1 compared to *FBN1* (9.7e-2 vs. 10.8e-2), NetREm ranks STAT1 higher as an activator for *FBN1* than for *APP* (13 vs. 20). This aligns with findings (Gu et. al 2022) of STAT1 LOF impacts of -0.61 for *FBN1*, -0.51 for *APP*.

NetREm reveals biologically relevant signals, identifying novel TGs for TFs. RXRG's high regulatory activity in nmSCs (**Table B.13** (SaniyaKhullar 2024)) is consistent with rodent/human studies (Gerber et al. 2021). All TGs across 3 groups (EGR2 mSC, SOX10 mSC and nmSC) are enriched in PNS-related terms (**Figure 3.3C**). We find 2,316 SOX10 TGs (159 shared, 428 direct, 1,314 LOFs overall) with 29 direct and 103 LOFs overall in both SCs. **Figure 3.3D** shows rat nerve epigenome tracks for 4 novel SOX10 candidate mSC TGs (not LOF or direct TGs of SOX10: *FAHD1, LARS2, SCAMP5, SOCS3*) with strong SOX10 binding to SC regulatory regions in open chromatin (Lopez-Anido et. al 2015) and **Figure B.15A-D** (SaniyaKhullar 2024) reveals the respective predicted locations of SOX10 binding to regions open chromatin in adult human SCs for these 4 TGs. In **Figure 3.3E**, SNP rs55927047 enhances TEAD1 binding to its TFBS on *FOXN2*'s promoter to help activate *FOXN2* in adult mSCs. Orthologous rat nerve TEAD1 ChIP-seq peaks also overlap with this promoter. This eSNP correlates strongly (is in linkage disequilibrium (LD)) with rs79073127 that links to higher inflammatory polyneuropathy risk in Pan UK Biobank GWAS (Turley et al). *FOXN2* GWAS-eQTL colocalizes for this condition, with 75% probability (Wallace 2020). Both SNPs correlate with increased *FOXN2* expression. We highlight additional TFs in SCs, underscoring their significant roles despite limited validation. Notably, our 297 mSC TGs for MEIS2 (core TF in DRG sensory neurons (Roussel et al. 2022)) are prominent in PPI pathways like: PI3K-Akt signaling (crucial for PNS myelination (Ishii et al. 2021)), actin cytoskeleton organization (essential for PNS regeneration (Wang et al. 2018b)) (**Figure 3.3F**). These TGs have higher

expression in pooled SCs compared to 4 CNS cell types (1-sided t, p-adj < 2e-6), a pattern absent in controls. **Figure B.16A** (SaniyaKhullar 2024) showcases principal hub TFs like EGR1 (485 nmSC eTGs) and RXRA (971 eTGs), regulating the most eQTL-validated eTGs in SCs. In **Figure B.16B** (SaniyaKhullar 2024), we present an example of an eQTL SNP linked to changes in the expression of TG *TTC3* in human nmSCs, potentially by altering YY1's ability to bind to TF binding sites along or near promoter proximal regions of *TTC3*. We found ChIP-seq peaks support YY1 binding to this orthologous region in rodents and noted that *TTC3* has strong GWAS-eQTL colocalization with nervous system-related disorders.

Figure 3.3 – Gene regulatory links between transcription factors (TFs) and target genes (TGs) in myelinating (mSCs) and non-myelinating (nmSCs) human Schwann Cells (SCs)

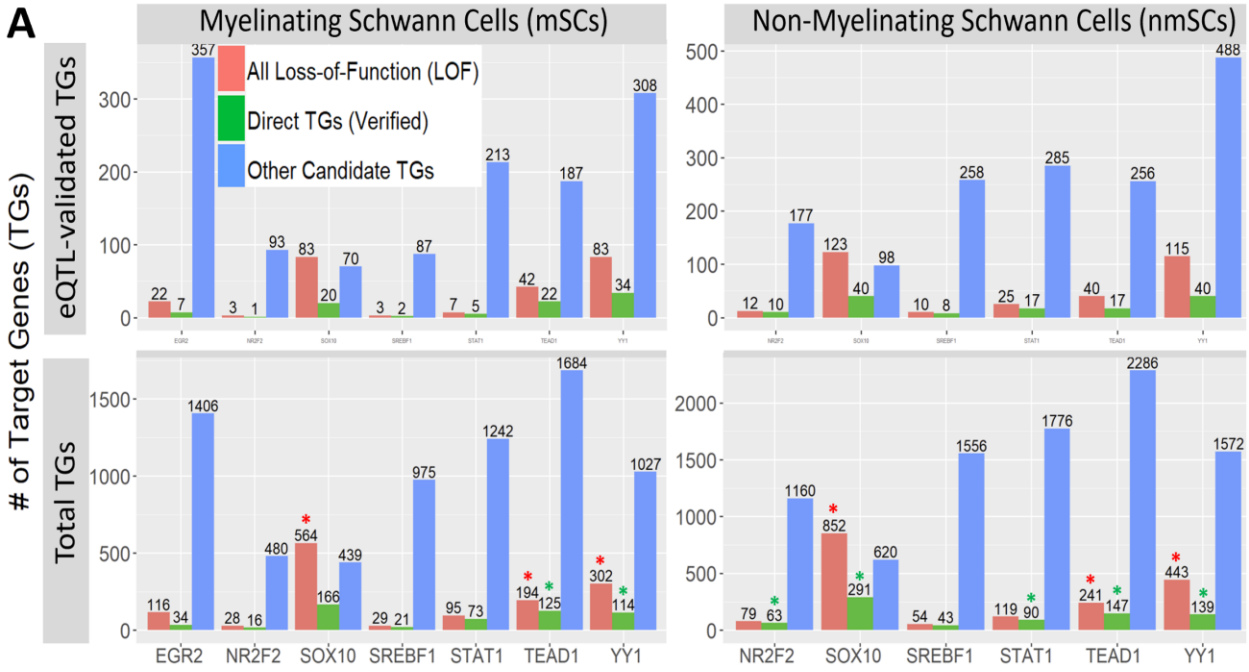


Figure 3.3A) - Bar plot categorizing TGs found for core TFs in mSCs (6 TFs: EGR2, NR2F2, SOX10, SREBF1, TEAD1, and YY1) and nmSCs (missing TF EGR2 given it is not a TF in the nmSCs). The top panel displays TGs with expression quantitative trait loci (eQTL) validation. The bottom panel reveals original counts with star (*) for over-enriched TGs based on hypergeometric test (p-adj < 5%).

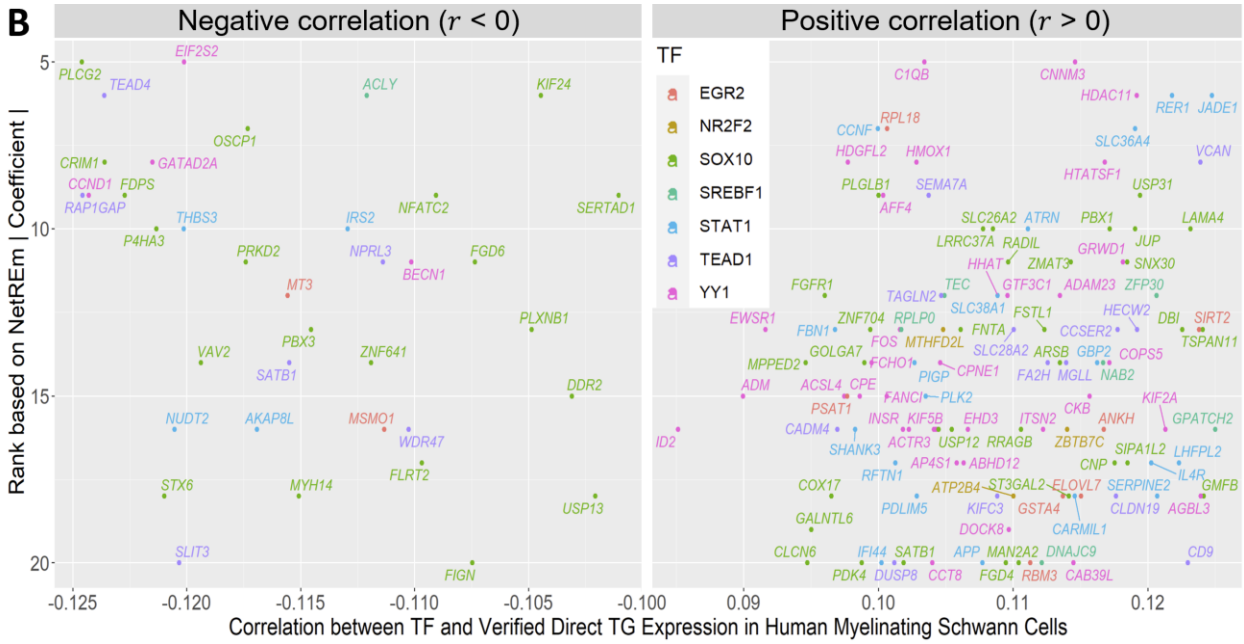


Figure 3.3B) -X-axis: correlation between core SC TF and validated direct TG (red: EGR2 TG, gold: NR2F2 TG, green: SOX10 TG, turquoise: SREBF1 TG, blue: STAT1 TG, purple: TEAD1 TG, pink: YY1 TG) based on training gene expression data for mSCs. Y-axis: rank of absolute value of NetREm's regression coefficient c^* for that TF for that given TG, where smaller rank values imply a greater magnitude for the coefficient (and stronger relationship), i.e. greater $|c^*|$. For simplicity, we show results where the TF is the top 5 to 20 for its direct TG.

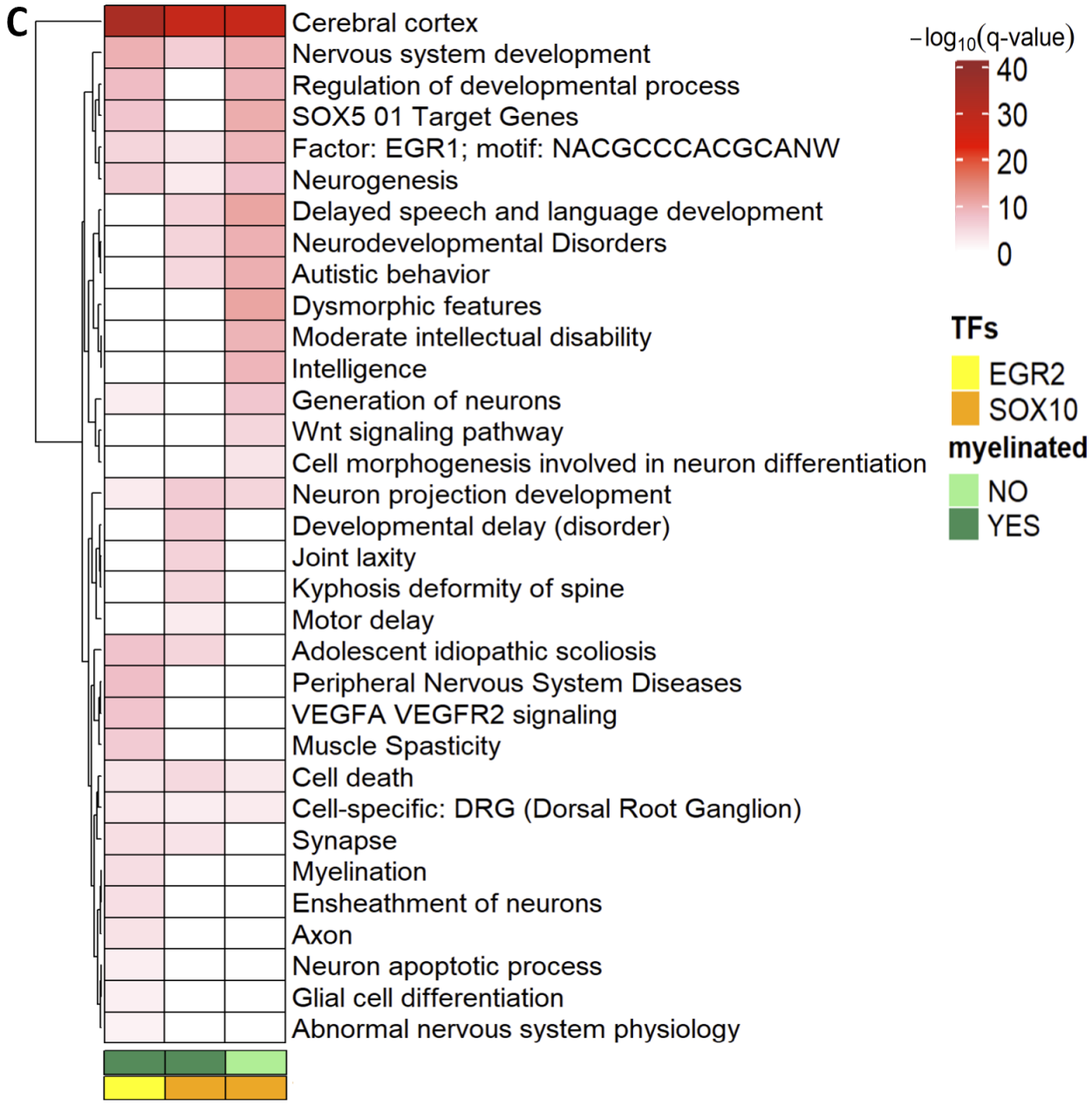


Figure 3.3C) - Select gene enrichments for all EGR2 mSC, SOX10 mSC, SOX10 nmSC TGs predicted by final NetREm model. Enrichments are hierarchically clustered, $-\log_{10}(q)$ values are reported.

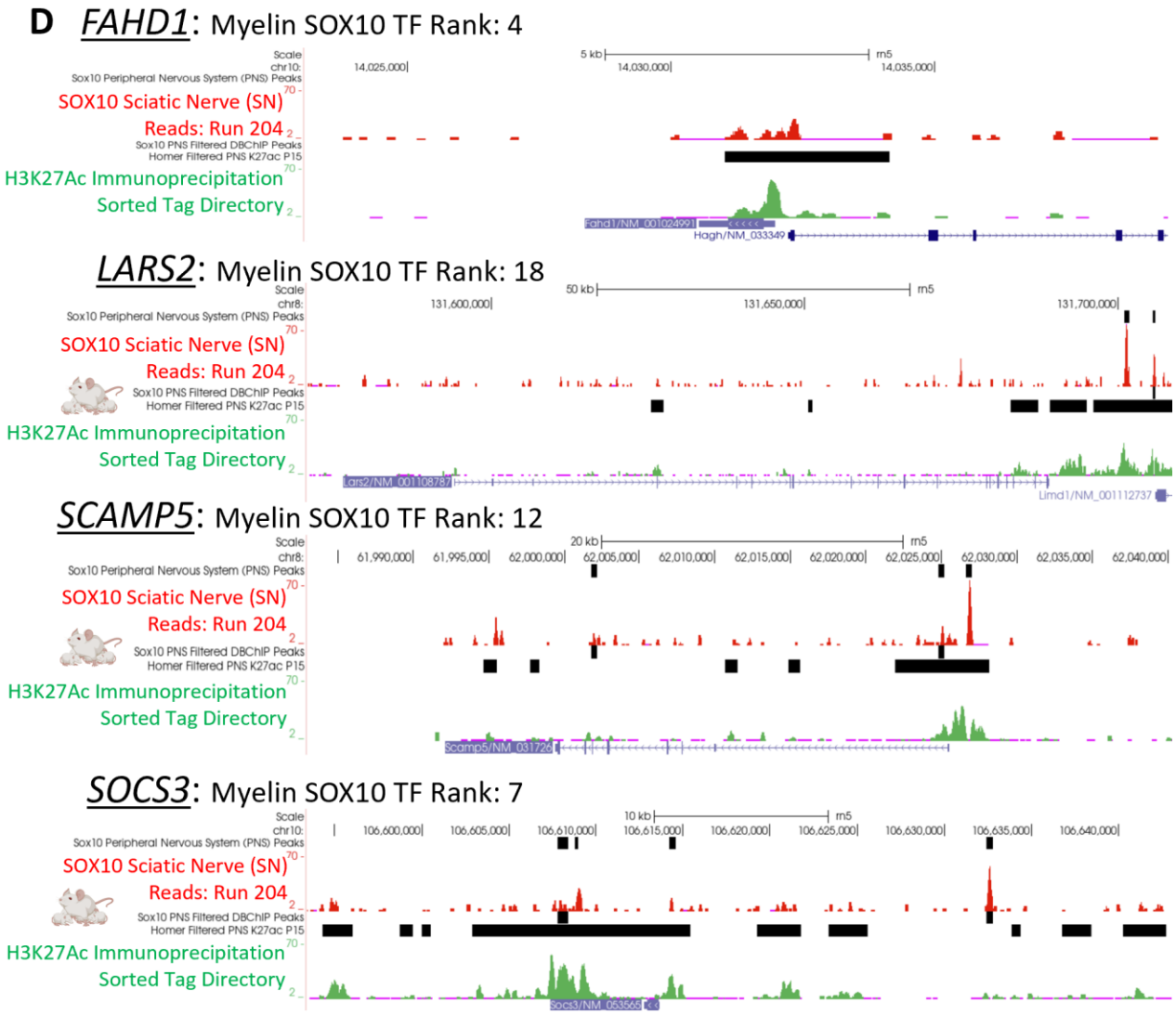


Figure 3.3D) - Epigenome tracks in rats (rn5 reference genome) for 4 potential novel candidate SOX10 mSC TGs with SOX10 as a top predictor TF. Here SOX10 is predicted as their activator TF (positive coefficient c^*). These tracks correspond to rat sciatic nerve ChIP-seq peaks, SOX10 peaks in the peripheral nervous system (PNS), histone modifications associated with enhancers (H3K27ac immunoprecipitation sorted tags).

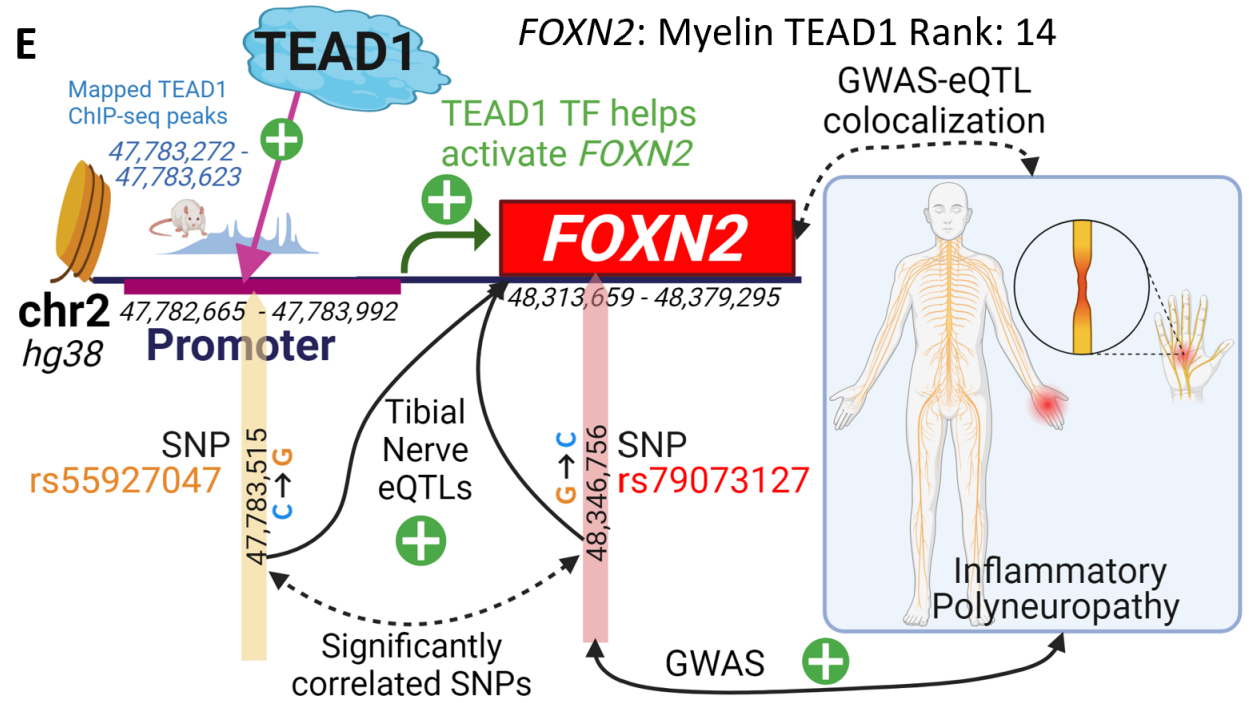


Figure 3.3E) - Tibial Nerve eQTL SNP (eSNP) rs55927047 (chromosome 2: position 47,783,515 in hg38 human reference genome, change in DNA base from C to G) located in the FOXN2 promoter (overlaps with orthologous TEAD1 ChIP-seq binding regions in rats) may strongly boost TEAD1 affinity for binding to the FOXN2 promoter to activate expression (given $c^* > 0$) of FOXN2, a candidate GWAS-eQTL colocalized TG biomarker for inflammatory polyneuropathy. FOXN2 is a novel TG for TEAD1 that has experimental support as a proximal TEAD1 TG based on ChIP-seq binding within 100 kilobases of its Transcription Start Site (TSS). Genome coordinates based on: hg38 human reference genome.

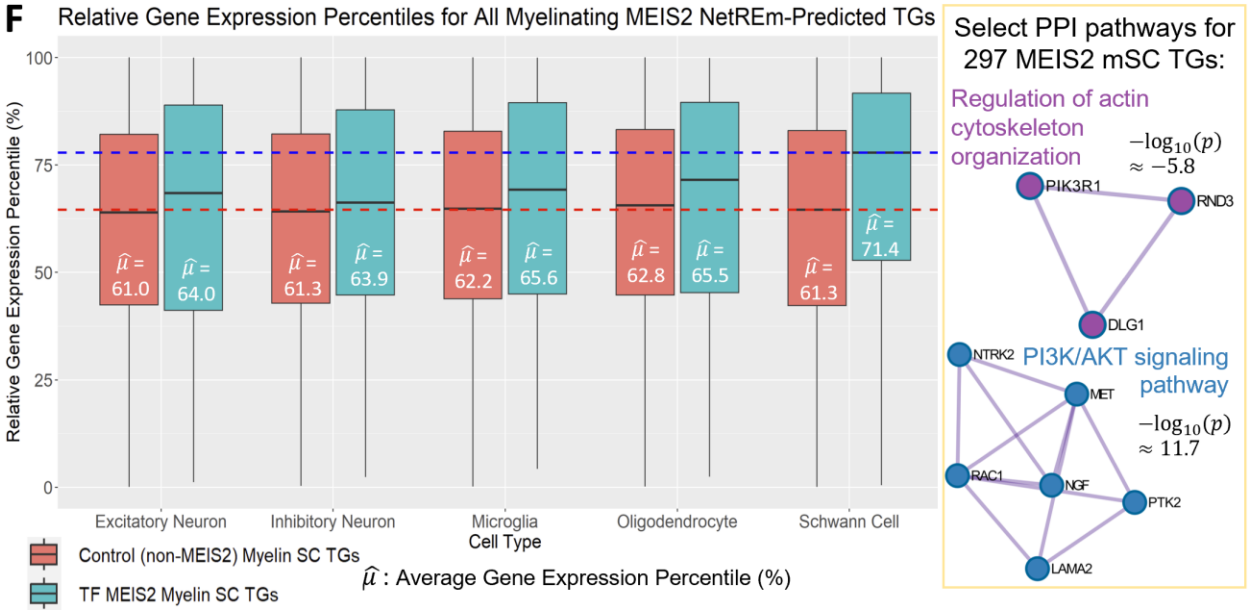


Figure 3.3F) - Left: Boxplots compare relative expression percentiles for all MEIS2 mSC TGs in GTEx pooled SCs (Eraslan et. al 2022)) with those in 4 CNS cell types (Microglia, Oligodendrocytes, Excitatory Neurons, Inhibitory Neurons (Lake et al. 2018)). Median percentile for MEIS2 mSC TGs in SCs overall is 71.4 versus ≤ 61.3 in controls (non-TGs for MEIS2 in mSCs). Right: enriched PPI paths for MEIS2 mSC TG proteins, which are important in SCs.

3.3.5 Coordination among transcription factors (TFs) for gene regulation in myelinating and non-myelinating human Schwann cells (SCs)

Our test MSEs are significantly lower than those of Linear and Ridge BRMs (**Figure 3.4A**, **Table B.14** (SaniyaKhullar 2024)). NetREm predicts SC-type-specific coordination B for each TG. It also outputs 22,809 mSC and 24,795 nmSC non-zero direct/indirect TF-TF coordination $\bar{B}$ links (**File B2** (SaniyaKhullar 2024)). Notably, top Context PPI Database (CPPID) contexts for strong mSC and/or nmSC $|\bar{B}|$ relate to the DRG region and the brain (**Figure B.17A** (SaniyaKhullar 2024)). **Figure 3.4B** shows 23 of 24 mSC-specific TFs in a mSC $\bar{B}$ network of 77 known PPIs, excluding novel links (w = 0.01) for simplicity (nmSCs: **Figure B.17B** (SaniyaKhullar 2024)); POU3F1-EGR2 mSC cooperativity is very strong ($\bar{B}$ = 99.14 percentile), JUNB-ATF4 ($\bar{B}$ = 96.98) interact in PNS neoplasms like Schwannomas. BNC2's regulatory activity in nmSCs may be attributed to the absence of its predicted repressor EGR2, ranked 10 of 20 mSC TFs for *BNC2*.

NetREm discovers and prioritizes novel TF-TF coordination links that are promising. 48 of our novel links in nmSCs (37 also in mSCs), comprising 30 TFs, are validated by strong SAINT scores of physical TF-TF binding in recent BioID/AP-MS human experiments (Göös et al. 2022) (**Figure B.17C** (SaniyaKhullar 2024)). RXRG, TEAD1, and YY1 co-regulate 366 nmSC (**Figure B.17D** (SaniyaKhullar 2024)) and 15 mSC TGs, suggesting their preferential coordination in nmSCs. These 4 core SC TFs co-regulate 174 TGs in nmSCs, *SETD2* in mSCs: RXRG, STAT1, TEAD1, and YY1 (**Figure B.17E** (SaniyaKhullar 2024)). In fact, RXRG links with STAT1, TEAD1, and YY1 are unknown in our input PPIN (**Figure B.18A** (SaniyaKhullar 2024)). RXRG strongly positively correlates with TEAD1 and YY1 (r = 0.5, 0.7) in nmSCs, but negatively in mSCs (-0.9, -0.7) in mice sciatic nerves (Gerber et al. 2021). STAT1 has 591 (Jaccard Similarity (JS): 0.21), YY1 has 599 (JS: 0.20), TEAD1 has 843 (JS: 0.26, significant) co-regulated nmSC TGs with RXRG (**Figure B.18B-C** (SaniyaKhullar 2024)). In nmSCs,

RXRG and YY1 share 94 eSNPs, 95 eTGs, 88 eQTLs compared with 28, 54, 24 for RXRG and STAT1 (cooperate for 704 TGs (**Figure B.18D** (SaniyaKhullar 2024)), antagonistic for 104 (**Figure B.18E** (SaniyaKhullar 2024))); RXRG- cooperation $\bar{B}$ is: TEAD1 (16.3), STAT1 (23.5), YY1 (24.1) (**Figure B.18F** (SaniyaKhullar 2024)).

To independently test our predicted TF-TF coordination for these core SC TFs, we use binding data from rat SCs in PNS, derived from ChIP-seq analysis of the active enhancer H3K27ac mark and ChIP-seq and CUT&RUN (Cleavage Under Targets and Release Using Nuclease) read density assay data of TF binding in nerve or S16 SC line. Of 15,864 ChIP-Seq H3K27 enhancer peaks shared between PNS and S16 lines, RXRG shares 3,450 and 2,017 peaks with YY1 and TEAD1 (**Figure B.18G** (SaniyaKhullar 2024)). 43.9% of RXRG peaks have YY1 binding (most colocalization among core SC TFs) and 25.7% of RXRG peaks have TEAD1 binding (**Figure B.18H** (SaniyaKhullar 2024)). Conversely, 24% of TEAD1 peaks and 28% of YY1 peaks have RXRG binding. CUT&RUN helps identify where TFs bind to DNA in the genome and determine the extent of colocalized binding along REs. TEAD1 CUT&RUN centered across EGR2 peaks reveals TEAD1 colocalizes at ~40% of EGR2 TFBSs (**Figure 3.4C**), supporting predicted EGR2-TEAD1 coordination in mSCs. YY1 CUT&RUN reads overlap ~70% when centered on RXRG peaks and SOX10 peaks. SC marker *CDH19* is preferentially expressed in nmSCs (Stratton et al. 2017) and has RXRG ChIP-seq binding nearby. Tibial Nerve eQTL SNP rs17799413 may be associated with lower *CDH19* expression (slope: 2.4e-1), strongly altering binding of TEAD1 in both SCs and 7 core TFs (e.g. RXRG, YY1) in nmSCs. CUT&RUN sequencing reveals TEAD1 and RXRG binding peaks are 40 kb upstream of *CDH19*'s TSS (**Figure 3.4D**). *NCAM1*, a YY1 LOF and SOX10 direct TG, codes for an adhesion molecule preferentially expressed in nmSCs (Martini and Schachner 1986; Wang et al. 2022a). Our network predicts regulation of *NCAM1* by TEAD1, YY1, and RXRG in nmSCs. Rs10749999 associates with higher *NCAM1* expression (slope: 9e-2) and may boost YY1 and TEAD1 binding in nmSCs. Active SC enhancers ≈130 and 200 kb upstream of *NCAM1*'s Transcription Start Site (TSS) have TEAD1 and YY1 binding. *NCAM1* has TEAD1, RXRG, and YY1 binding at a promoter surrounding its gene locus. Rat H3K27Ac and TF ChIP-seq data shows RXRG co-regulates: 555 TGs

with TEAD1, 352 with YY1, and 27 of 91 mapped STAT1 TGs. This further supports our predicted coordination by these 4 TFs, showing NetREm's prowess in identifying novel colocalizing TFs in the absence of current evidence of direct binding interactions from high-throughput studies of PPIs.

Computational methods help decipher functional impacts of SNPs on PPIs, aiding in uncovering disease risk genes for targeted precision therapies. Most non-coding, disease-associated SNPs alter human PPIs rather than protein properties like folding or stability (Cheng et al. 2021). Integrating TF-TG regulatory (c^*) and TF-TF coordination (B) networks with non-coding eSNP rs11663049 sheds insights on how PPIs associate with phenotype: polyneuropathy in diabetes (**Figure 3.4E**). Dysregulated mitotic checkpoint regulators may lead to abnormal insulin signaling in diabetes (Choi et al. 2016). *CEP192* helps form mitotic spindles (Joukov et al. 2010) and colocalizes for this phenotype with 69% probability. The eSNP reduces the risk of this phenotype and *CEP192* expression and may strongly decrease binding of 6 activator TFs to *CEP192*'s promoter in nmSCs. *CEP192* is co-regulated by RXRG-STAT1 TG. RXRG-NR4A2, NR4A2-THRB associate with PNS neoplasms. SOX10, with its dynamic, cell-type-specific cooperation, works with these TFs; SOX family TFs achieve cell-type-specificity via partner TFs that facilitate TG regulation by binding to nearby SOX TFBSs (Stevanovic et al. 2021). SOX10-NFIA cooperate in nmSCs ($\bar{B}$ = 81.3) but display antagonism influencing glial lineage diversification (Glasgow et al. 2014) in mSCs ($\bar{B}$ = -78.7).

We present a similar example (**Figure B.19** (SaniyaKhullar 2024)) for mSCs, where we predict how SNP rs9847953 is associated with altered regulation of TG *ZNF589* in mSCs, by potentially impacting a TF-TF coordination network comprising 5 TFs (boosting binding and/or TFBS motif affinity of 3 activators and decreasing that for 2 repressor); overall, this SNP-TG pair is significantly associated with injuries to the nervous system, in particular: injury of nerves at wrist and hand levels. While *ZNF589* is a final TG in our mSC regulatory network, it is not found as a TG in our final nmSC regulatory network.

To illustrate differences in TG-specific TF-TF coordination networks B between SC sub-types, even for shared TGs, we examine TG *ART3* that colocalizes with neurofibromatosis (NF: characterized by

the formation of NS-related tumors) with a 68% probability (**Figure 3.4F**). We predict an interacting DNA chromatin loop (Zhang et al. 2021) of REs in open chromatin in adult SCs, featuring differing $\bar{B}$ between mSCs and nmSCs involving strong repressors and activators (Sharov et al. 2022) to regulate *ART3*. 3 eSNPs link to lower *ART3* expression: NF-associated rs4859594 correlates with 2 regulatory SNPs (rs6856681: intronic enhancer; rs9998233: promoter proximal RE) that may disrupt coordination networks by strongly decreasing activator binding to *ART3's* REs, increasing it for repressors. Both SCs types have SOX2 as a common TF; it is a core regulator of SC myelination and myelinating disorders, and a super pioneer TF that previous studies note is associated with cancer cell proliferation and survival (Benedetti et al. 2022). SOX2 coordinates with TFs at *ART3's* proximal promoter, eagerly cooperating with: TEAD1 (B = 15.76) in nmSCs, SETDB1 (B = 24.2) in mSCs. PPARA-NFIX (nmSCs) and NR3C2-PBX3 (mSCs) are antagonistic relations on the enhancer.

Figure 3.4 – Coordination among transcription factors (TFs) for gene regulation in myelinating (mSCs) and non-myelinating (nmSCs) human Schwann cells (SCs)

Figure 3.4A) - Density boxplots: NetREm outperforms Linear Regression and RidgeCV in both SCs, with lower test MSEs. NetREm predicts links for more TGs than ElasticNet and Lasso do. For mSCs, NetREm achieves median MSE: 0.95; nmSC: 1.1. We focus on the same TGs in each SC sub-type (based on final predictions): 3,248 in mSCs and 5,980 TGs in nmSCs.

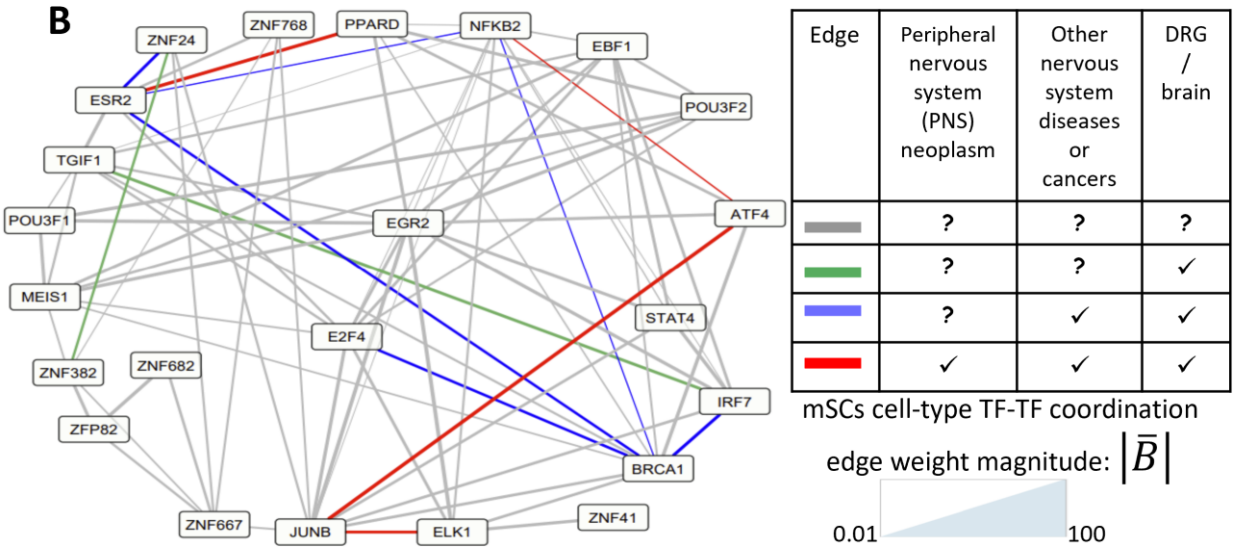


Figure 3.4B) - Input TF-TF PPIN subnetwork for only known TF-TF PPIs among 23 of 24 mSC-specific TFs only; artificial (novel) links excluded. Edges represent cell-type TF-TF cooperation ($\bar{B}$) across TGs in mSCs in the mSC NetREm TF-TG Regulatory Network, since $\bar{B} > 0$ for all. High $\bar{B}$ = stronger cooperation in mSCs. Edges annotated based on Contextual PPI database (Kotlyar et. al 2022).

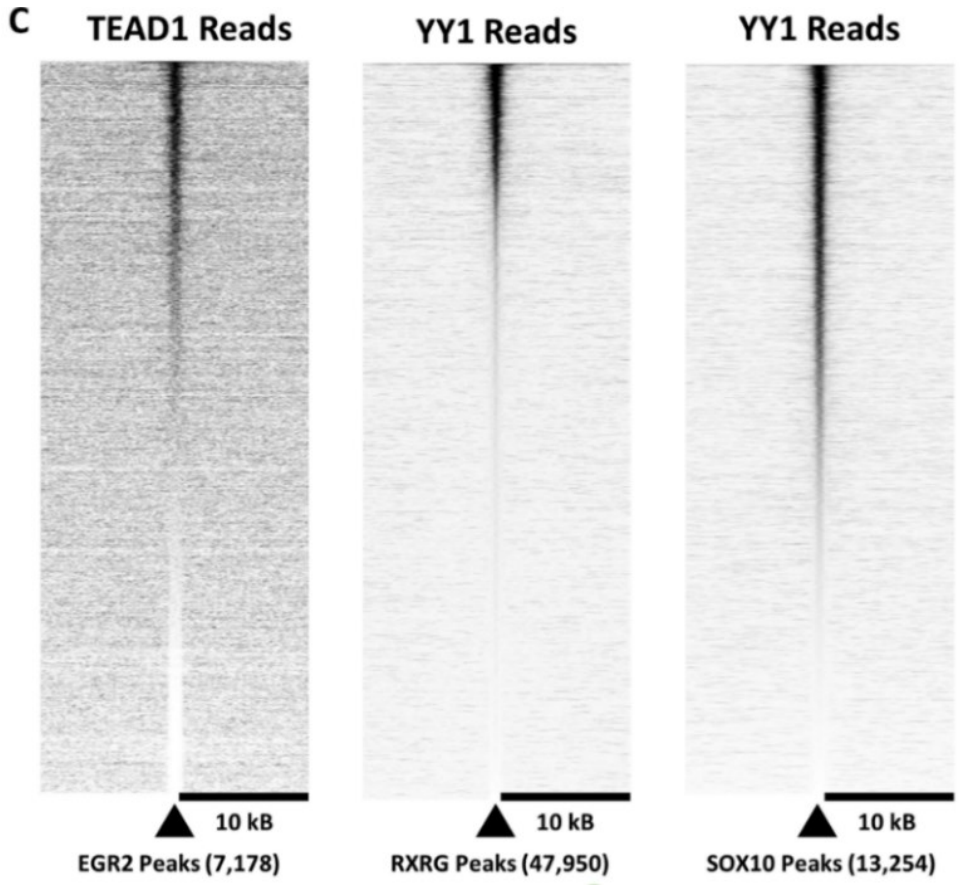


Figure 3.4C) - Left to right: Heatmaps show read density overlaps of the TEAD1 CUT&RUN assay centered on EGR2 peaks, YY1 CUT&RUN read density centered on RXRG peaks, and YY1 CUT&RUN read density centered on SOX10 peaks.

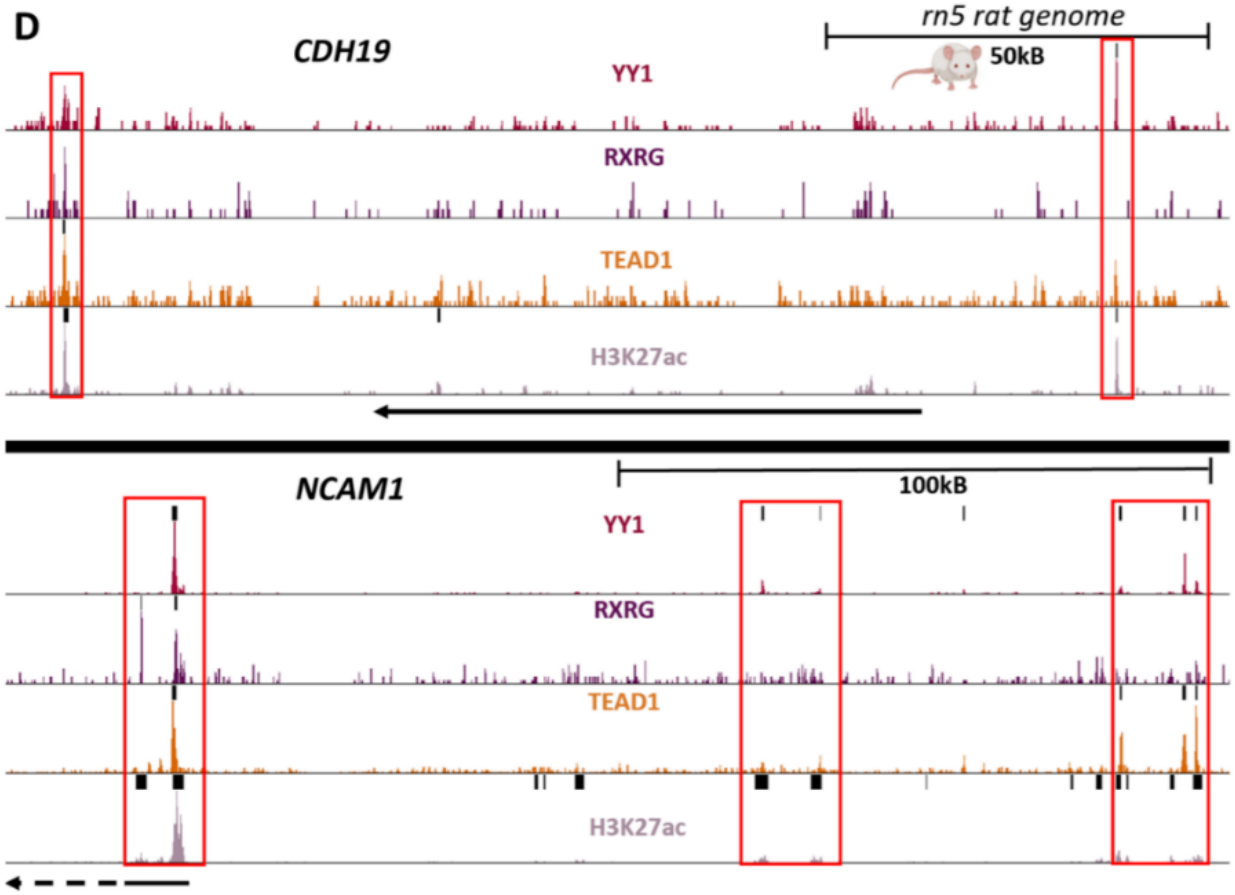


Figure 3.4D) - CUT&RUN sequencing tracks of YY1, RXRG, and TEAD1 TFs, along with ChIP-seq tracks of H3K27ac enhancer peaks, are shown from the S16 Schwann cell line in rats. Boxes highlight enhancer regions where TFs colocalize across genes *CDH19* (top tracks) and *NCAM1* (bottom tracks). Both TGs are impacted by Tibial Nerve eQTL SNPs (eSNPs) that alter regulatory TF binding at their loci.

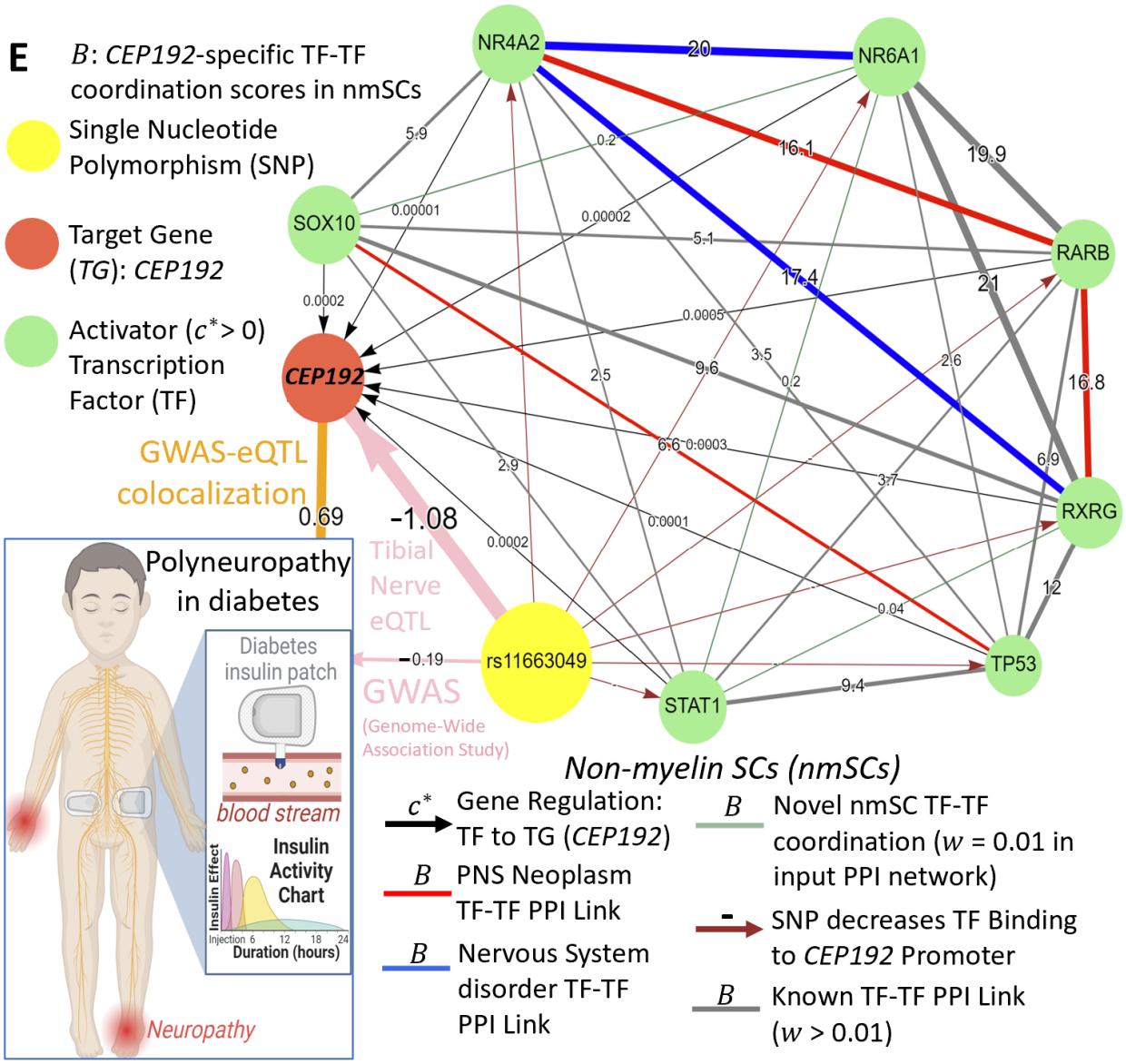


Figure 3.4E) - While **Figure 3.3D** focuses on the impact of a single eSNP on 1 TF, this panel provides a more extensive example for TG *CEP192* in nmSCs, which GWAS-eQTL colocalizes with decreased risk of polyneuropathy in diabetes. Tibial Nerve eSNP (yellow node) is associated with decreased risk of this condition and decreased *CEP192* expression by disrupting bindings (brown arrows) of 6 activator TFs (positive coefficients $c^* > 0$, green nodes) and boosting bindings (black arrows) of potential repressor TFs ($c^* < 0$, blue nodes) to a *CEP192* regulatory region (i.e. promoter). SOX10 cooperates with them. *CEP192*-specific raw TF-TF *B* coordination scores are undirected links. Functionally-validated direct and/or indirect TF-TF interactions (found in the input PPI network) are grey links and novel TF-TF links (not in input PPI network) are colored teal. Further, these functionally-validated TF-TF interactions are colored if they are associated with PNS Neoplasms (red) or Nervous System-related disorders (blue) based on the Contextual PPI database (Kotlyar et. al 2022).

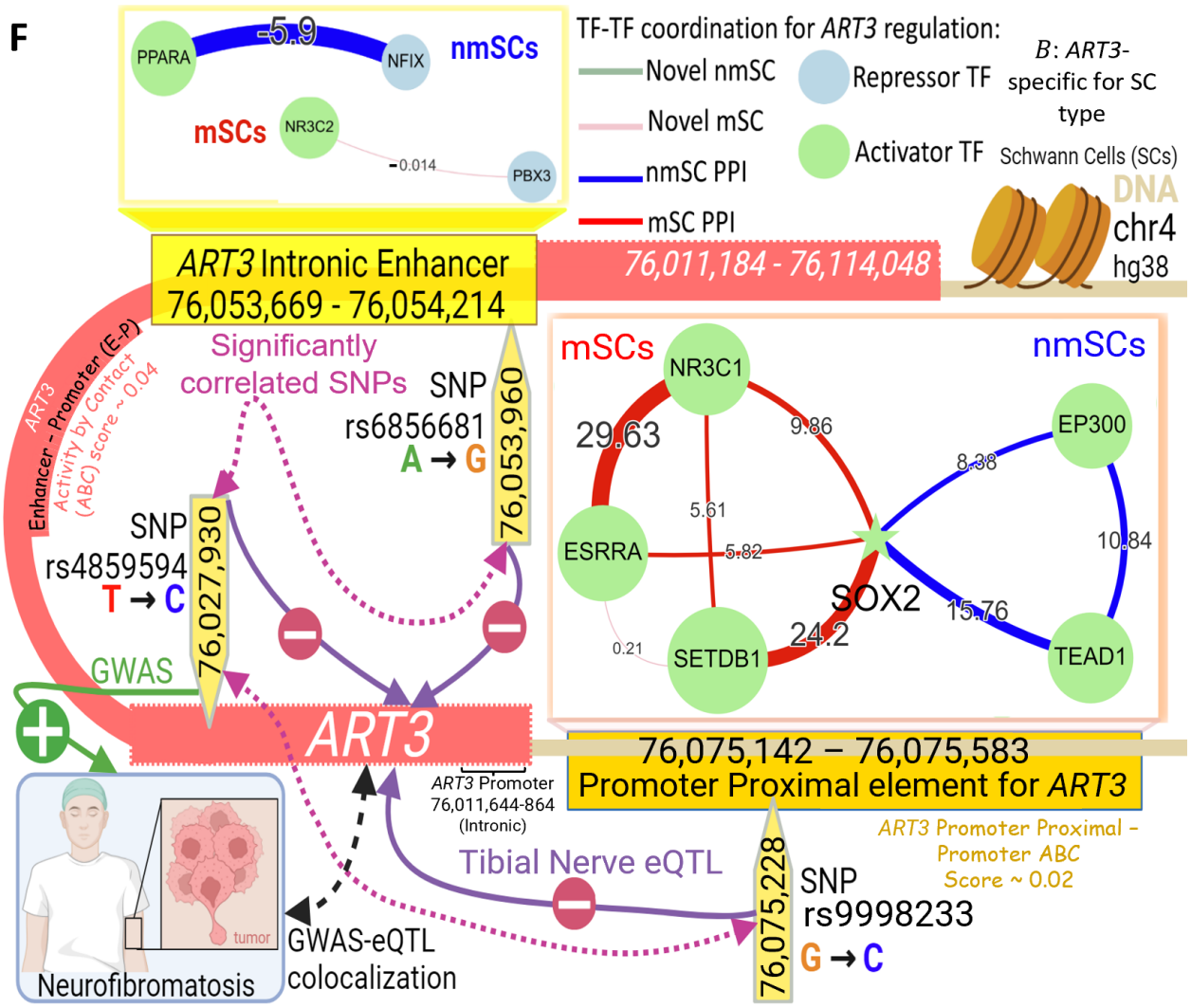


Figure 3.4F) - Tibial Nerve eSNPs potentially influence regulation of GWAS-eQTL colocalized neurofibromatosis TG *ART3* through activators and repressors forming distinct TRMs in mSCs and nmSCs along interacting REs (regulatory elements) in a 3D loop of open chromatin in adult SCs. ART3's (intronic) promoter has Activity by Contact (ABC) scores (Zhang et al. 2021) of ~0.04 with intronic enhancer, ~0.02 with proximal promoter RE. TF-TF *B* edge widths are shown relative to other TF-TF *B* links in that given *ART3* RE (i.e. intronic enhancer, promoter proximal). Blue and red links refer to direct and/or indirect TF-TF interactions in nmSCs and mSCs, respectively, with support from the input PPI; the corresponding pink and green links are for the remaining TF-TF coordination links that are novel (and may constitute direct and/or indirect TF-TF relationships, which follow-up studies may investigate).

3.3.6 Prediction & comparative analysis of cell-type coordination among TFs for gene regulation across neuronal/glial cell types in Alzheimer's disease

Cell-type-specific TF-TF coordination networks $\bar{B}$ are crucial for neuronal functions like synapse plasticity and neurotransmission, and are disrupted in AD, leading to memory loss, neuroinflammation, cognitive decline (Mathew et al. 2022). Understanding how altered $\bar{B}$ impacts TG expression in AD is

essential for identifying master regulators and developing targeted therapies (Wang et al. 2016). Signaling PPIs associated with dementia symptoms highlight the potential of targeting these altered PPIs (Mao et al. 2020) to delay AD progression (Vargas et al. 2018).

In response, we integrate multi-omics data to construct 16 context-specific prior GRNs for 8 cell types in AD and Controls. For each TG in a context, NetREm uses the N TG-specific candidate TFs from the respective prior GRN as input features, based on TFs that may associate with its REs to regulate it (**Methods**). Ultimately, NetREm generates 16 corresponding TF-RE-TG (TF-TG links: **File B3** (SaniyaKhullar 2024)) and $\bar{B}$ (**File B4**, word cloud: **Figure B.20A** (SaniyaKhullar 2024)) networks. Reverse engineering changes in networks across cell types may help illuminate molecular drivers of AD.

We evaluate and quantify the scale-free characteristics of our predicted TF-TG regulatory networks by comparing NetREm with scNET (Gupta et al. 2022), which applies the scGRNom pipeline (Jin et al. 2021) to generate TF-RE-TG networks based on the same preprocessed, underlying gene expression data for 8 contexts: 4 out of 8 cell types in AD versus Control stages. Across all 8 contexts, NetREm demonstrates superior scale-free topology, with power law degree exponent (γ) higher than scNET's γ by between 0.343 and 1.032 and with coefficient of determination (R^2) higher than scNET's R^2 by between 0.031 and 0.421 (**Table B.15** (SaniyaKhullar 2024)). These improvements ranges over scNET indicate that NetREm more accurately captures properties of real-world biological networks, with γ closer to the ideal 2-3 range and R^2 nearer to 1 (Chen et. al 2023; Broido and Clauset 2019; Langfelder and Horvath 2008).

To assess biological relevance, we used the comprehensive and recently developed brainSCOPE resource (Emani et al. 2024; brainSCOPE Resource 2024), which provides a high-resolution atlas of cell-type-specific GRNs with detailed transcriptional profiles, as a proxy for signed ground truth networks in Controls. We compared TF-sign-TG regulatory links predicted by NetREm and scNET in Control Microglia (Mic) and Oligodendrocytes (Oli), carefully filtering the respective networks to retain only genes common to NetREm, scNET, and brainSCOPE, followed by an additional filter for brainSCOPE-identified cell-type-specific TFs and TGs. NetREm outperforms scNET, achieving a higher Jaccard similarity (JS) score of 0.083 in Mic, a 59.6% improvement over scNET's 0.052, and 0.053 in Oli, a

23.3% increase over scNET's 0.043. We also computed weighted signed average Area Under Precision Recall (AUPR) metrics on these filtered networks (**§B.1** (SaniyaKhullar 2024)), ensuring that AUPR calculations focus on valid and comparable transcriptional interactions across all models. Overall, NetREm outperforms scNET in predicting transcriptional regulatory interactions when compared to the brainSCOPE baseline in both Control Mic and Control Oli. In Mic, NetREm achieves an average AUPR of 0.687, compared to 0.645 for scNET, while in Oli, NetREm similarly performs better with an average AUPR of 0.713 versus 0.648 for scNET. These results reinforce NetREm's stronger biological relevance and predictive accuracy, highlighting its robustness in predicting meaningful TF-TG regulatory interactions across different cell types and disease contexts.

We explored TG-specific coordination B in Control vs. AD stages for 2 AD risk genes (Jia et al. 2020; Bossaerts et al. 2022) (**Figure 3.5A**). *TMPRSS15* in Mic (t: 36.1, p-adj < 4.8e-283) and *ABCB5* in Inhibitory Neurons (InNs) (t: 22.6, p-adj < 4.5e-112) show notable increases in B in AD versus baseline Controls for the respective cell-type. Some TF-TF pairs strongly cooperate ($50 \leq B \leq 100$) exclusively in AD: ZBTB14-ZNF281, FLI1-TAL1 for *TMPRSS15*; ZNF331-ZNF354A, MYEF2-SOX2 for *ABCB5*. For *TMPRSS15*, known AD links IRF7-STAT3 and STAT3-STAT5B have strong antagonism (i.e. $-100 \leq B \leq -50$) in controls but cooperate in AD. **Figure 3.5B** compares strong $\bar{B}$ among select TFs between conditions for Mic and InNs. IRF7-STAT3 is in Mic, MYEF2-SOX2 is in InNs only. AD link STAT3-STAT4 is antagonistic in Controls for *TMPRSS15* in Mic and in Control Mic overall (but cooperative for other 3 networks). RORA-ELK1 is cooperative in all 3 but antagonistic in Control Mic. FOSL2-BACH1 is antagonistic in Mic but cooperative in InNs. Indeed, RORA activity increases in InNs and Mic in AD (Acquaah-Mensah et al. 2015). AD-annotated links in AD InNs/Mic include ELK1-SPI1 and ELK1-STAT3.

Figure 3.5C presents a multifaceted network weaving together TF-TG regulatory links, TF-TF coordination, phenotypes, SNPs. TF-TF coordination is discernibly stronger and positive in Control Mic (1-sided t), pointing to potentially disrupted cooperativity during AD. Our attention is drawn to regulation of *ANXA11*, a critical player in diverse functions (e.g. apoptosis, neutrophil function) and signaling paths

(e.g. MAPK, P53) (Mirsaeidi et al. 2016). *ANXA11* is 1 of our 36 AD-Covid genes (**Chapter §2, Table A.8**). Mutations in *ANXA11* are correlated with NS diseases (Wang et al. 2022b) and high risk of inflammatory conditions like sarcoidosis (Smith et al. 2017). Our AD-Covid study (Khullar and Wang 2023) assigns *ANXA11* to a Hippocampal Control gene co-expression module. NetREm offers nuanced insights on non-coding SNPs for AD (rs11202929) and cognitive function (rs12412257), linked to more and less *ANXA11* expression, respectively, in Mic (resident CNS macrophage immune cells). Rs12412257 associates with reduced word interpolation ability, a measure of fluid intelligence and reasoning (Turley et al) and prognostic marker of AD (Eyigoz et al. 2020). Rs11202929 is protective against AD (GWAS slope < 0) and may enhance binding affinity of 12 cooperating activators to TFBSs on *ANXA11's* enhancer in open chromatin in adult Mic. SPI1 is a core Mic TF for AD genes (Rustenhoven et al. 2018b). *ANXA11* has higher Mic expression in controls than in AD (t p-adj = 2.6e-68, $\log_2$(Fold Change) of (Control/AD) = 0.78; 668 AD, 676 controls). Our significant GWAS-eQTL colocalization analyses reveals that *ANXA11* expression positively associates with lower AD risk and with better word interpolation ability. NetREm provides a powerful framework to deepen our understanding of complex GRNs and their implications across a spectrum of health conditions.

We further evaluate our $\bar{B}$. **Figure B.20B** (SaniyaKhullar 2024) compares $\bar{B}$ across 16 networks for 216 novel links validated in recent physical human experiments (Göös et al. 2022). In control ExNs and InNs, $r \approx 0.54$ and 0.45, respectively, between $\bar{B}$ and Jaccard Similarity (JS) of ChIP-seq peak overlap (a metric used as a proxy for cooperativity (Yu et al. 2015)) for 6 TFs in neural cells (**Figure B.21A-C** (SaniyaKhullar 2024)).

For each cell-type, we build default Random Forest (RF) (**Figure 3.5D)**, Logistic Regression, Naïve Bayes, and XGBoost machine learning models (Pedregosa et al. 2011) to detect TGs with altered B from Controls to AD that may predict TF-TF links annotated (in CPPID) with neurodegenerative disease (**Figure B.22A-C** (SaniyaKhullar 2024)). Input data consists of changes in B across TGs from Control to AD, for the given cell type. To tackle this positive unlabeled learning problem (Yang et al. 2014), we

undersample class 0 so each cell-type has an equal count of TF-TF links in both classes (**§B.1** (SaniyaKhullar 2024)). We evaluate performance via stratified 5-fold CV noting RF is the most optimal approach as all 8 RF cell-type models have average area under Receiver-Operator Characteristic (ROC) curve (AUC) ≥ 0.81. Across our models, the top 500 RF feature TGs (in terms of feature importance scores) are enriched for neurodegeneration, cell-type functions, immunity, intellectual disabilities, tauopathy (**Figure 3.5E**). OPCs and InNs have the highest overlap (35 top TGs) (**Table B.16** (SaniyaKhullar 2024)); disrupted InN signaling to OPCs may diminish myelination and CNS interneuron activity, and severely impair prefrontal cortical network functions and social cognitive behavior (Fang et al. 2022).

Figure 3.5 – Prediction and comparative analysis of cell-type coordination among transcription factors (TFs) for target gene (TG) regulation across neuronal and glial cell types in Alzheimer's disease (AD)

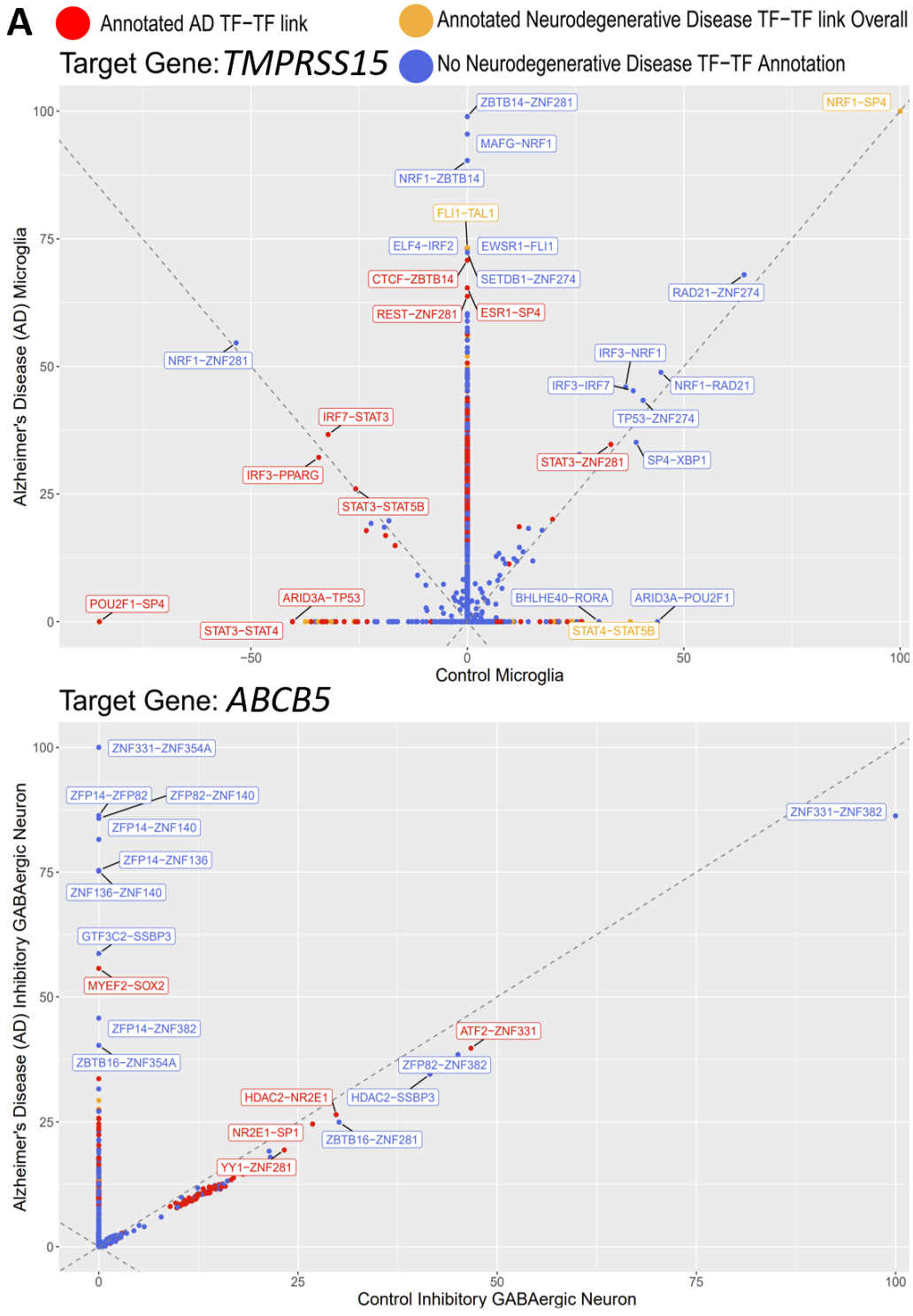


Figure 3.5A) - Scatterplots comparing raw TF-TF links and corresponding coordination scores *B* for 2 AD risk TGs, *TMPRSS15* for Microglia and *ABCB5* for Inhibitory Neurons, in Control versus AD. Red and orange points are TF-TF links annotated by Contextual PPI database (CPPID) (Kotlyar et. al 2022) for AD and other neurodegenerative diseases, respectively. Blue points: no annotation for neurodegeneration.

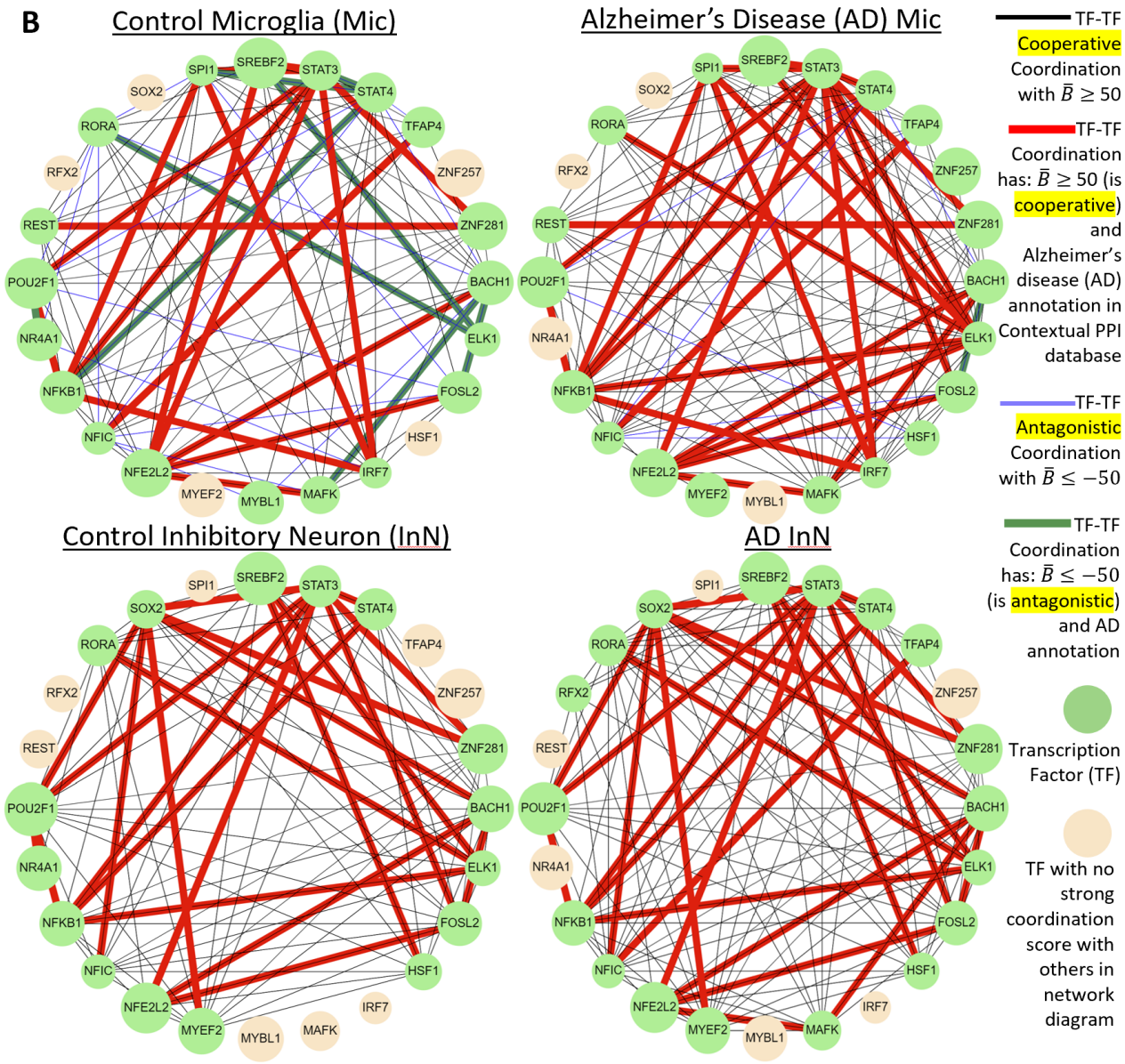


Figure 3.5B) - Circular network diagrams visualizing cell-type TF-TF coordination (direct and/or indirect interactions) in AD/control Microglia (Mic) and Inhibitory Neurons (InNs), focusing on select TFs denoted by light green nodes. These TFs exhibit strong and potentially cooperation ($50 \leq \bar{B} \leq 100$) or antagonism ($-100 \leq \bar{B} \leq -50$). Red links are AD-related in the Contextual PPI database (CPPID)(Kotlyar et. al 2022). TFs not expressed or lacking strong links in this filtered network are in peach. Thick edges are AD-related in CPPID. Peach: TFs not expressed or lacking strong links with other TFs in this visual.

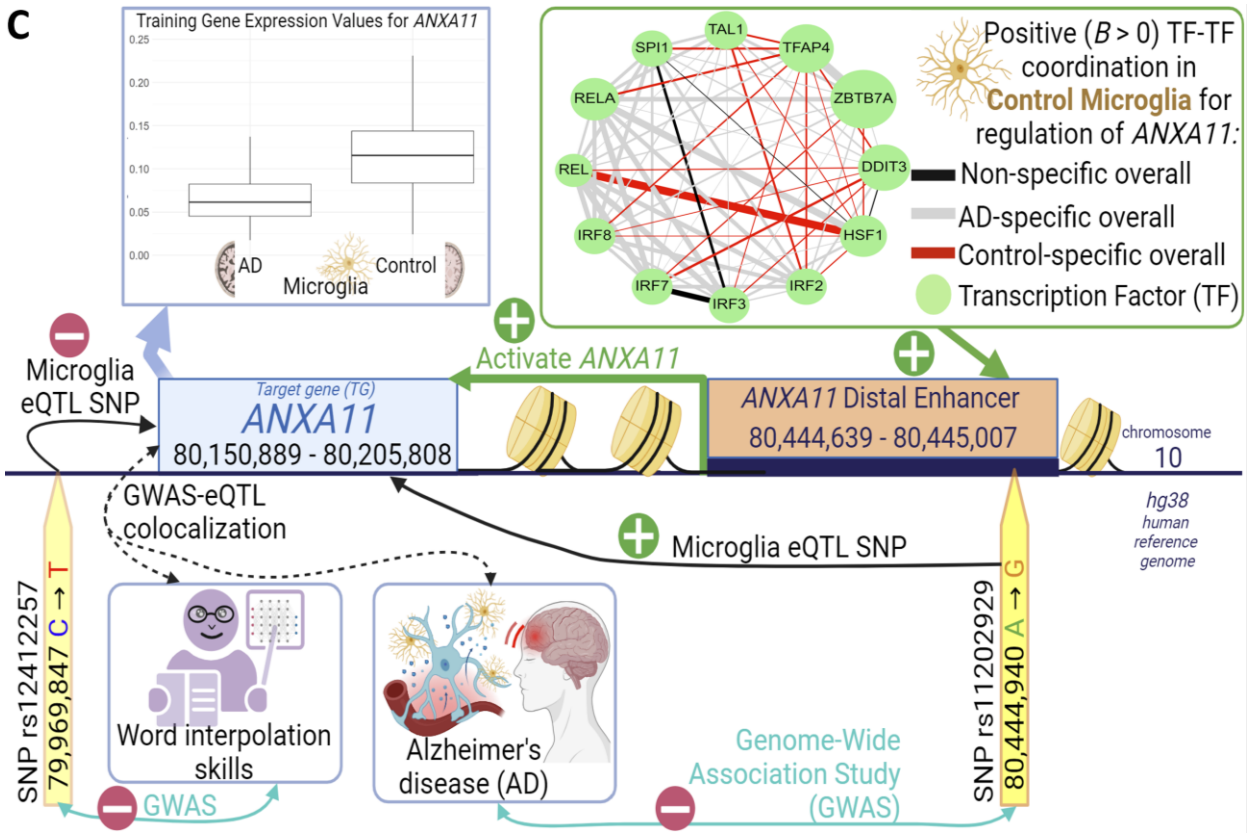


Figure 3.5C) - 12 TFs that may cooperate to activate *ANXA11* in TF-TG regulatory network in Control Mic (higher expression for *ANXA11* than in AD). Widths for TF-TF links are *B* scores for *ANXA11* regulation in Controls and colors are based on statistical significance of TF-TF links across all TGs in both Mic networks. Light gray links: higher in AD; red: higher in Controls; black: not significant. A SNP correlating with lower AD susceptibility, increases binding of this TF-TF network, and links to higher *ANXA11* expression. Risk SNP for reduced cognitive ability (poor word interpolation skills) links to lower *ANXA11* expression.

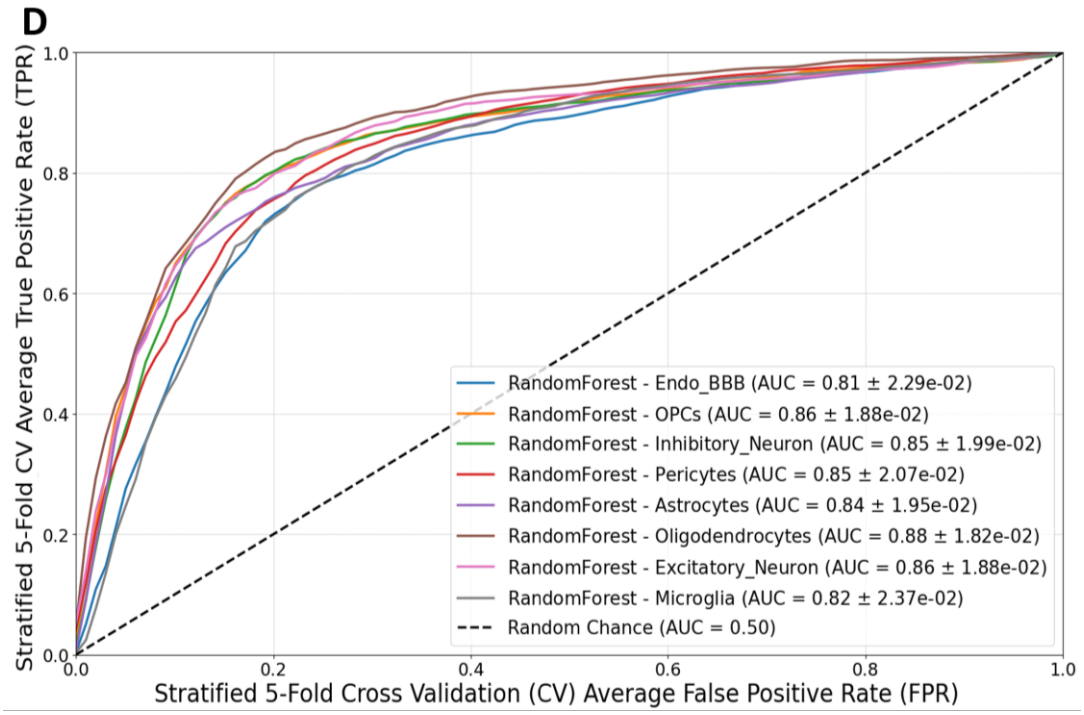


Figure 3.5D) - Average Stratified 5-fold Cross Validation (CV) Receiver Operator Characteristic (ROC) curves for Random Forest (RF) models designed to predict TF-TF links annotated in neurodegenerative diseases (class 1) in the Contextual PPI database (CPPID) (Kotlyar et. al 2022) based on balanced class data (equal # of links in class 0 and class 1).

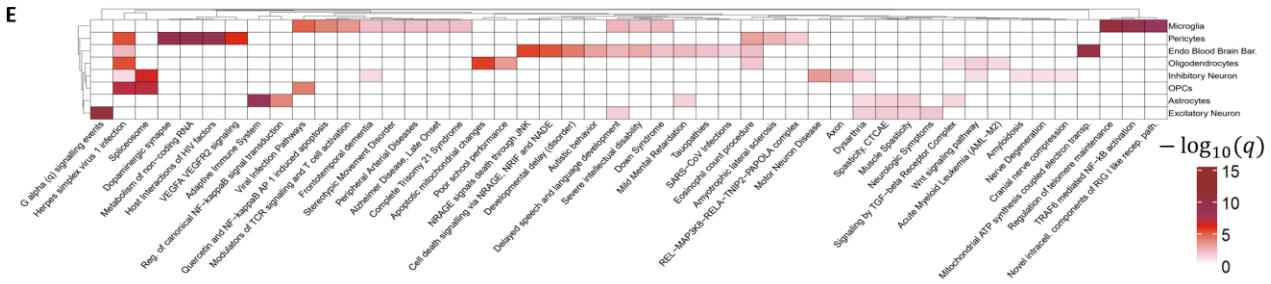


Figure 3.5E) - Heatmap of gene enrichment analysis terms for 500 optimal genes (top feature importance) for each cell-type identified by Random Forest (RF) models for 8 cell types. Hierarchical clustering is performed on rows (cell types) and columns.

§ 3.4 Discussion

In this chapter, we present NetREm, a computational multi-omics-based approach that employs network-regularized regression on single-cell gene expression data to predict cell-type coordination among Transcription Factors (TFs) for target gene (TG) regulation. NetREm addresses a major challenge in traditional studies of cell-type gene regulatory networks (GRNs). Gene expression data, often nascent,

sparse and noisy, fails to capture crucial GRN mechanisms such as TF binding to DNA, coordination among TFs/cofactors, and DNA accessibility, and typically provides weak signals for distinguishing TP from FP TF-TG links (Kim et al. 2023; Badia-i-Mompel et al. 2023). Sole reliance on expression data for GRN inference is therefore woefully inadequate, and perhaps even futile, often leading to unstable and inaccurate results.

Explicitly modeling direct and indirect TF-TF interactions can enhance GRN inference, enabling discovery of novel TF-TG links and key cell-type TFs (Skok Gibbs et al. 2022). Functionally-related TF predictors, like neighbors in scale-free feature networks (e.g. TF-TF PPINs), can coordinate synergistically or antagonistically in biological processes like TG regulation (Kong and Yu 2018). Nonetheless, traditional methods often miss such complex dynamics of TF-TF PPIs involved in GRNs (Yazaki et al. 2016).

State-of-the-art cell-type GRN inference tools like SCENIC indirectly hint at TF-TF interactions by analyzing TFs that co-regulate multiple TGs. However, these tools primarily focus on TFs with strong motif binding, excluding other prior information. For instance, in our comparative analysis, SCENIC identifies many TFs in SCs (nmSCs: 640, mSCs: 522) but overlooks TEAD1 in both SCs, instead detecting 3 other TEAD family TFs. TFs like TEAD1, which exhibit relatively weak motif-binding signals, are drowned out. In contrast, by incorporating TF-TF PPINs, NetREm effectively captures essential GRN relations for core SC TFs like TEAD1. This underscores the importance of integrating comprehensive prior information like PPINs, in GRN predictions from expression regression, a capability that NetREm successfully implements (Dibaeinia and Sinha 2020).

NetREm reveals cell-type coordination among TFs, $\bar{B}$, with some mediated by physical and others by indirect (e.g. pioneer/settler models show TFBSs are often >50 bp apart in REs) PPIs (Martin et al. 2023). By weighing known direct/indirect PPIs in the context of TG regulation, NetREm helps characterize existing PPINs at a cell-type level (Johnson et al. 2021; Murtaza et al. 2022; Hsu et al. 2022). It also helps address the link prediction problem, flagging undiscovered PPIs for follow-ups (Singh and Vig 2017). The lack of cell-type PPI annotations, while a challenge, offers NetREm an opportunity to

contribute to ongoing efforts to broaden understanding of PPIs and protein dynamics with its dual capacity to annotate known TF-TF PPIs and discover novel cell-type-specific ones (Yu et al. 2023).

NetREm predicts unprecedented, weighted, cell-type-specific TF-TF coordination networks $\bar{B}$ across various conditions, including both human and mouse contexts, even in the absence of prior GRNs. Our benchmarking showed our TF-TG regulatory networks not only performed competitively with SOTA GRNs but also uncovered novel cell-type TFs that coregulate TGs; further, our $\bar{B}$ effectively prioritized TP and FN TF-TF links.

Disrupted cell-type PPIs are critical in neurobiological disorders, since PPIs mediate neuronal functions (Mathew et al. 2022). We integrated multi-omics data, capturing various levels of TG regulation (e.g., scRNA-seq, scATAC-seq), to learn prior GRNs for NS cell-types for our SC and AD human applications. Detecting these candidate GRN TFs for TGs is key to inferring biologically significant cell-type TF-TG links (Zaborowski and Walther 2020; Zhang et al. 2023). Aligning NetREm's TF-TG links with prior GRNs helps us deduce TF-RE-TG links. NetREm uncovered novel TF-TF crosstalk for TG regulation in SCs and during AD in neurons/glia.

Insights derived from NetREm may contribute to advancing targeted therapies and regenerative medicine. We apply our predicted regulatory and coordination networks to trace how non-coding eSNPs may alter co-regulatory dynamics among TFs, potentially altering expression of disease-associated eTGs.

NetREm expands upon previous work in network regularized regression by learning and generating embeddings using SVD. In the future, we can incorporate nonlinear dimensionality reduction into NetREm to capture nonlinear patterns and regularize latent representations with prior information. We may see if any final TFs form homodimers or adapt NetREm to account for this. Additionally, regularization networks can integrate other information, like signaling pathways, to learn TF-TF coordination for TG regulation.

Beyond expression regression, we can extend NetREm to other emerging single-cell omics like scATAC-seq to explore TF and chromatin interactions in open regions. In **Table B.17** (SaniyaKhullar 2024), we provide additional examples of potential applications of NetREm to the field of biology (e.g.

disease gene identification, drug response prediction, epistatic interactions among SNPs that influence complex traits/disease, calculate individual polygenic risk scores considering genetic variant interactions). NetREm extends to any discipline where predictors exhibit a network structure that informs the outcome.

§ 3.5 Availability of data, software, and materials

We implement NetREm as an open-source software package: GitHub.com/SaniyaKhullar/NetREm with details in **§B.1** (SaniyaKhullar 2024). **Tables B.18 - B.19** (SaniyaKhullar 2024) provide a breakdown of the # of TGs, # of TFs if N is fixed (applications 1 – 5) or metrics if variable (applications 6 – 7). For instance, in hESCs, we run NetREm 1,250 times (1 for each TG), with N = 206-207 TFs for each TG. **Table B.20** (SaniyaKhullar 2024) lists resources, data, and materials utilized. We use human hg38 and rat rn5 reference genomes. Please note that metrics for NetREm's final cell-type outputs (TF-TG gene regulatory networks (GRNs) and TF-TF coordination networks) in applications 6 (human SC sub-types: mSCs versus nmSCs) and 7 (AD versus control across 8 neuronal/glial cell types) are available in **Table B.21** with corresponding **Figure B.23**. Our corresponding methods and materials, figures (**Figures B.1-B.23**), tables (**Tables B.1-B.21**), and data files (**Files B1-B4**) for NetREm are available in **§ Chapter B** of the supplementary file (SaniyaKhullar 2024) that is hosted at: https://github.com/SaniyaKhullar/Supplementary_Chapters_Dissertation.

§ Chapter 4: Conclusion and future work

§ 4.1 Summary, limitations, and future work for SNPheno

4.1.1 Summary of SNPheno

In **§ Chapter 2**, we applied SNPheno to perform an integrative multi-omics study to predict AD GRNs along with gene co-expression modules for three major brain regions. Using these networks and modules, we further linked several AD–Covid genes that improve AD and severe Covid predictions (Zhang et al. 2024), and also revealed regulatory mechanisms of genome-wide association study (GWAS) SNPs of AD and of severe Covid-19.

4.1.2 Limitations of SNPheno, potential alternatives, and future work

SNPheno has several potential pitfalls that we have considered and tried to limit. There are also limitless avenues for future work, which we are currently pursuing.

4.1.2.1 Exploring other avenues whereby non-coding SNPs may impact TG expression

The first is that our work only considered gene regulation from TFs, but disease variants can impact other regulatory mechanisms (e.g. histone modifications and DNA methylation) to further effect gene expression. In addition to SNPs, we could consider other genetic variants like insertions/deletions or structural variants (copy number, etc.) or RNA-binding protein variants. Nonetheless, researchers are generating population-level epigenetic and whole genome sequence data to systematically identify epigenetic activities and structural variants. Such data can help improve our predicted GRNs. There are upcoming methods for multi-omics profiling of single cells, including CITE-seq, SLIDE-seq, paired-seq, single-nucleus chromatin accessibility (Sealfon et al. 2021).

4.1.2.2 Extending SNPheno to Single-cell gene expression data for matched individuals

Our application of SNPheno is in terms of bulk-level data for 3 brain regions. In the future, there will be epigenomics data available for individuals with various disease conditions (e.g. Hi-C chromatin interaction data for individuals with AD at the brain region and/or cell-type-level), which we can utilize to build more precise cell-type and/or brain region GRNs for individuals in various conditions. Moreover, multi-omics data may be disparate and unmatched among individuals in the population, which may result

in inconsistencies; nonetheless, there are upcoming data resources for multi-omics data for matched individuals. There are recent single-cell gene expression data sets (e.g. (Mathys et. al 2019)) for Alzheimer's disease versus controls, which SNPheno may be applied on. To build gene co-expression networks (where genes with similar expression dynamic patterns are likely co-regulated and assigned to common or similar gene co-expression module), we used WGCNA. This approach is typically used for bulk-level gene expression data since single-cell gene expression data (e.g. scRNA-seq or snRNA-seq) is inherently sparse. Our pipeline may be extended to single-cell data by recent methods such as scWGCNA (Morabito et al. 2021), to identify disease-associated and other phenotype-associated modules of co-expressed genes at the cell-type level. In addition, we can run scGRNom tool to infer cell-type reference gene regulatory networks (GRNs), by using recent human data (e.g. (Zhang et al. 2021)) on chromatin interactions among cell-type candidate cis-regulatory elements (cCREs) in adult and/or fetal stages.

4.1.2.3 Integrating eQTLs to improve SNPheno networks

Another limitation is that many disease loci identified by GWAS studies do not coincide with any known eQTL. A potential reason why is due to cis-bulk-eQTL data. In future, we can integrate cell-type-eQTLs and trans-eQTLs (along with cis-eQTLs) for novel insights into trans-regulatory mechanisms associated with diseases. More data is being generated on cell-type eQTLs, which we can integrate in the future to validate our findings. There is emerging single-cell data that can help uncover more disease risk genes and details of transcriptional regulation at much finer resolutions. In response, there are large-scale international collaborative efforts, like the single-cell eQTLGen consortium (van der Wijst et al. 2020), which is collecting a large sample size of data in individual immune cell types to better understand cell-type specific effects in cis and trans-genes, and prioritize risk genes. For instance, scQTLbase (Ding et al. 2023) is a recent comprehensive database and visualization platform combining 304 datasets, across 57 cell types and 95 cell states. Further, the recent brainSCOPE (Emani et al. 2024) resource provides cell-type-specific eQTLs from single-cell data for several nervous system-related brain cell types.

Currently, we use independent eQTLs to help prioritize and validate SNPheno links from non-coding SNPs to impacted TGs. Given the recent availability of cell-type-specific eQTLs from various resources, we can use some resources to directly improve SNPheno networks instead. To reduce the potential for False Positive (FP) links, we may extend SNPheno to utilize eQTL data and GWAS-eQTL colocalization analysis to filter SNP-effected-GRN predictions that are made and improve SNPheno networks overall. In the future work, we are extending SNPheno to build cell-type gene regulatory networks for thousands of traits based on GWAS studies like the Pan UK Biobank.

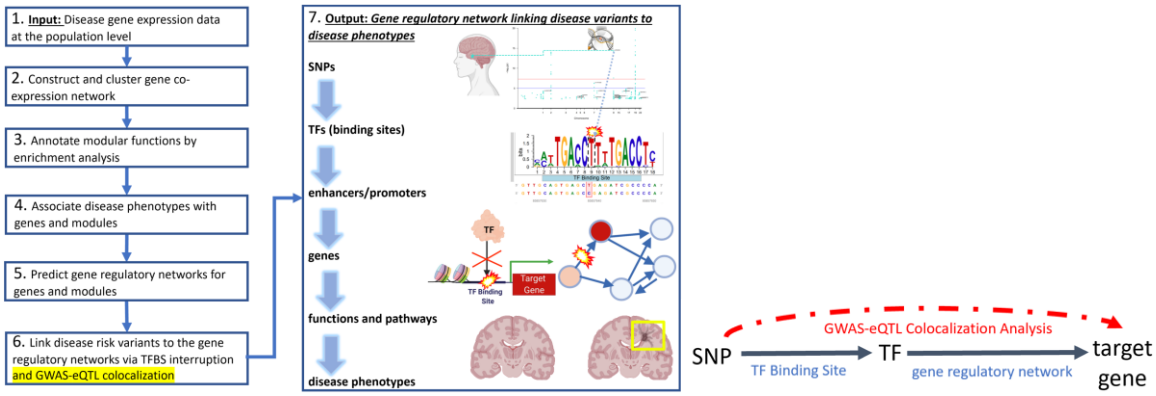


Figure 4.1 – Adapting the SNPheno pipeline to incorporate GWAS-eQTL colocalization

We update the pipeline for SNPheno (**Figure 2.1**). We add GWAS-eQTL colocalization to Step 6. Colocalization analysis reveals links between SNP and TG expression for more robust results.

Further, we may predict the impact of insertions/deletions (i.e. indels) of DNA bases on TF binding. Currently, SNPheno focuses on the impact of SNPs (1 DNA variation) on TF Binding. Perhaps we could predict how deletions of regions coinciding with TFBSs for TFs may disrupt TF binding and regulation of that respective TG; similarly, insertions of regions coinciding with TFBSs for TFs may lead to potentially new binding of TFs for the given TG.

4.1.2.4 Incorporate Protein-Protein Interaction (PPI) networks into SNP-effected-GRNs

We can potentially incorporate PPI networks to better understand how translational machinery may be involved in disease risks (central dogma of molecular biology: gene to protein (via transcription) to protein (via translation). Thus, we can integrate PPI networks into our SNP-effected-GRNs to determine groups of TFs that interact together (e.g. cooperativity (+) or antagonistically (-)) or validate gene co-

expression modules based on findings of their protein products interacting more often with each other. Studies have observed a community structure within PPI networks (Padi and Quackenbush 2015), with groups of proteins interacting together with each other, to associate with biological functions and processes. These studies find that disease genes are often located nearby in the human PPI and there may be disease modules of PPIs implicated in diseases. In this way, we may extend SNPheno (that is at a transcriptional level) to predict post-translational effects and modules. Thus, we may associate the GRNs and gene co-expression modules in our SNP-effected-GRNs with PPIs, to prioritize disease modules and biomarkers at the protein level (e.g. disease-associated PPI modules). We may combine SNPheno with NetREm (incorporates PPI networks as a constraint to learn TF-TG regulatory networks) to create more holistic GRN links that also include coordination among TFs.

4.1.2.5 Heterogeneous Graph embedding techniques to reveal relationships among SNPs, predict gene regulatory links, and functionally annotate roles of SNPs

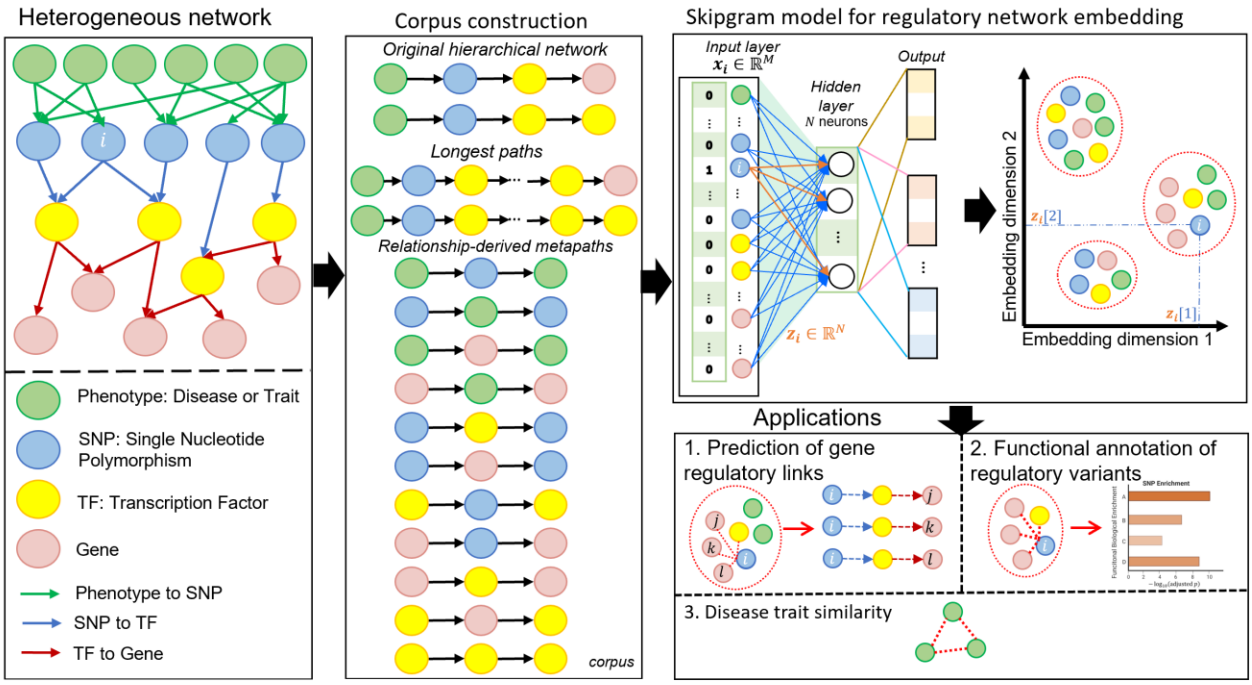


Figure 4.2 – Embedding SNPheno heterogeneous networks to uncover novel relationships among nodes and predict new links

SNPheno builds heterogeneous graphs linking non-coding phenotype-associated SNPs to TGs they may help dysregulate (via altered gene regulatory network mechanisms). We note different relationships and

various nodes (e.g. phenotype, SNP, genes: TF versus target gene) and edge relationships. These networks are hierarchical. We can learn relationships among these nodes by graph embedding approaches. Then, we perform graph-embedding to reduce and meaningfully summarize the graph's dimensionality using a variety of potential techniques (that preserve higher order graph proximities among nodes) like: DeepWalk (Perozzi et al. 2014), node2vec (Grover and Leskovec 2016), Graph Convolutional Networks. For instance, node2vec uses a Skipgram (Mikolov et al. 2013) neural network model on a corpus of node sentences, to learn embeddings. Node2Vec is a powerful algorithm designed for learning continuous feature representations for nodes in a graph, useful in various machine learning applications such as node classification and link prediction. It leverages a biased random walk strategy to explore the graph, capturing both the local and global network structures. By balancing depth-first and breadth-first sampling through parameters p and q, Node2Vec generates a diverse set of node sequences, which are then fed into the Skipgram model, originally developed for natural language processing tasks. The Skipgram model processes these sequences to maximize the likelihood of preserving node neighborhoods in the embedding space, producing low-dimensional vectors that reflect the complex structural and relational patterns of the original graph. This approach allows Node2Vec to create embeddings that are highly informative and applicable to a wide range of network analysis tasks.

These embeddings can be used in various downstream tasks. We can perform link prediction tasks to predict effects of rare or never-seen SNPs, or SNPs whose impact has not yet been observed but may soon be observed through more technological advances in sequencing. We may look at other methods by which non-coding genetic variants impact gene expression (besides TF binding disruption), like chromatin modifications and DNA accessibility. Biology is complex so we can utilize link prediction approaches on network graphs (shown to be more accurate using graph embeddings than other data types (Goyal and Ferrara 2018)) to predict missing interactions or potential links that SNPheno has not yet identified. Link prediction tasks can be invaluable towards predicting effects of rare or never-seen SNPs, or SNPs whose impact may soon be observed (through technological advances in sequencing (Wong et. al 2021)). We could build SNPheno networks using different GWAS datasets for a given disease and see if

link prediction methods for older GWAS SNPheno networks predict links observed using newer GWAS (assuming newer GWAS is conducted on a larger sample size and returns more meaningful results).

We can use SNPheno to help annotate the potential function of clusters of SNPs. For instance, we may perform enrichment of SNPs based on enrichment of genes in their cluster or by aggregating GWAS traits that the SNPs are significant for. In addition, we can use cosine similarity among phenotype nodes to uncover a network that reveals potential phenotype trait similarity.

4.1.2.6 Adjusting the epigenomic data to account for dynamic DNA accessibility

SNPheno uses epigenomic data to determine regions of euchromatin (open chromatin) on DNA for TFs to recognize their sequence-specific motifs to bind to help regulate expression of their TGs. Thus, SNPheno does not consider the ability of TFs to bind to heterochromatin (closed or compact chromatin). Nonetheless, there are pioneer TFs, special TFs that can recognize their TFBSs on heterochromatin, which triggers remodeling of the chromatin landscape to make it more accessible for other TFs (i.e. non-pioneer TFs) to bind (Mayran and Drouin 2018). Moreover, DNA accessibility can fluctuate dynamically (Klemm et. al, 2019) and considering chromatin accessibility as binary (open or closed) may be harmful for GRN prediction (Miao and Kim 2022). It may be helpful to adjust the underlying GRN methods to consider DNA accessibility more quantitatively (Badia-i-Mompel et al. 2023). In the future, we may incorporate freshly-available epigenomic data on chromatin accessibility in disease states to infer more accurate disease-specific GRNs.

§ 4.2 Summary, limitations, and future work for NetREm

4.2.1 Summary of NetREm

NetREm helps highlight the intricate interplay of TFs in regulating TGs across various cellular contexts. The complex coordination among TFs, crucial for TG expression, remains largely uncharted. Traditional models mostly focus on individual TFs, overlooking the collaborative dynamics vital for TG regulation. NetREm innovates beyond these limits, offering a refined, network-regularized regression model that unravels sophisticated pathways of direct/indirect TF coordination. Our model assimilates multi-omics

data (e.g. PPIs, epigenomic markers, TF binding, chromatin interaction), fostering comprehensive networks for TF-TG regulation and TF-TF coordination.

4.2.2 Limitations of NetREm, potential alternatives, and future work

There are limitations to NetREm and areas for further research and innovation.

4.2.2.1 Incorporating negative network weights in network-regularized regression model

Covariates in regression models can be interconnected on a network, and incorporating this network structure into regularized regression models via a network penalty term can improve model performance. In the future, we may extend NetREm to incorporate any negative interactions that may be known in the input PPIN. Currently, NetREm assumes that the input TF-TF PPIN has positive weights $w > 0$. This is since the PPI represents the strength and probability of 2 protein nodes interacting and does not focus on the nature of their interaction. That is, whether proteins act antagonistically (-) or cooperatively (+) in a pathway, their corresponding PPI weight in current PPI networks simply denotes the magnitude of their interaction. Nonetheless, there are on-going efforts to annotate the nature of PPIs. In the future, some PPIs could be annotated as being cooperative or antagonistic. Hence, input TF-TF PPINs may eventually contain negative weights as well as positive ones. Given this, we may refer to network regularized regression studies for insights and inspiration on modeling not only positive but also negative weights. For instance, the iterative algorithm called 3CoSE (3-step Covariate and Connection Sign Estimation) (Weber et al.) has been developed. When connection signs (+ or -) in the network are unknown, they must be estimated alongside covariate coefficients. This 3CoSE algorithm alternates between estimating the connection signs (e.g. +: activating, cooperating; -: repressing, antagonistic) and covariate coefficients. Simulation results and an application in forecasting event times demonstrate the algorithm's effectiveness across various settings.

4.2.2.2 Incorporating self-weights in the input network weight matrix W

Most TFs cooperate with other TFs (rather than operate in isolation), working in concert to regulate TG expression utilizing mechanisms like co-binding or tethered-binding (Nie et al. 2020). TFs can be part of stable complexes (Ibarra et. al 2020) or they can enhance binding affinity of other TFs to nearby TF

binding sites (TFBSs) (synergistic activation) to regulate TGs (Ibarra et. al 2020; Zhao 2023). It is not uncommon for the regulation of one TG to necessitate interactions with 10-15 TFBSs (Bentsen et. al 2022). Many TFs are unable to bind alone and instead typically require PPIs to form homomeric (TF pairs with identical TF: $TF_i - TF_i$ links) or heteromeric (pairs with different TFs: $TF_i - TF_j$ links) complexes prior to binding to DNA (Morgunova and Taipale 2017). Despite the prevalence of such TF-TF coordination, the underlying intricacies of this phenomena are not yet fully comprehended (Ibarra et. al 2020). NetREm currently uncovers different TFs that coordinate to regulate a given TG in a particular cell-type and context. In our regression problem, we do not consider self-loops (i.e. TF interactions with itself); instead, we design our W so that the main diagonal represents a given node's average connectivity to all $N - 1$ other nodes in the TG-based input PPI network excluding itself (i.e. its average degree d_i). In the future, we may extend NetREm incorporate functionality to model homodimer PPIs. Eventually, as a second step, NetREm can utilize known homodimer PPIs to determine whether any of those final TFs forms a homodimer. Another alternative is to adjust the NetREm model to incorporate self-loops to explicitly model homodimers. To do this, we will have to reformulate the W matrix.

4.2.2.3 Non-linear dimension reduction to reveal embeddings

NetREm learns gene expression embeddings ($\tilde{X}, \tilde{y}$) using singular value decomposition (SVD). In the future, we can integrate nonlinear dimensionality reduction into NetREm to reveal nonlinear and prior knowledge regularized latent representations.

4.2.2.4 Different regularization approaches to predict TG expression from gene expression embeddings

We utilize Lasso regression to predict gene expression embedding values of our target gene (TG): $\tilde{y}$ from the gene expression embeddings of the N candidate TF predictors $\tilde{X}$. We provide an alternative ElasticNetREm implementation that uses ElasticNet($\tilde{X}, \tilde{y}$) regression instead of Lasso($\tilde{X}, \tilde{y}$) regression. That is, we adapt our Network regularized regression model to utilize other regularization models based on the gene embeddings from network regression (besides Lasso) such as ElasticNet. Currently, we use Lasso regression. In the future, we could explore Ridge regression models to solve this problem. That is,

we can input the gene embedding regressions into problems involving other types of regularizations such as L2 norm. We can add more regularization terms and explore the effects.

4.2.2.5 Non-linear objective function to predict TG expression from embeddings

In the future, we may consider other non-linear objective functions to predict $\tilde{y}$ from $\tilde{X}$. Studies (e.g. (Wang et al. 2023)) observe that mutual information (MI)-based models may perform better than linear based models at detecting GRNs since TFs often exhibit nonlinear behavior (e.g. cooperativity, oligomerization: formation of protein complexes to regulate gene expression). We may explore utilizing the gene regression embeddings (combining prior gene regulatory knowledge, gene expression data, and PPI weights) as input data not just for linear regularization models, but as input for Mutual Information (MI)-based models. We may also consider PoLoBag (Polynomial Lasso Bagging) (Roy et. al 2020) on our gene embeddings for signed GRN inference (activator vs. repressor) as this tool may help incorporate not only linear relationships (1^{st} order polynomial interactions) but also non-linear relationships (through higher-order multiplicative interactions). We may consider an ensemble approach (e.g. TreNA Ensemble Solver (Arment S et al. 2021)) to annotate TF-TG regulatory links based on different machine learning models that predict that respective link (e.g. Lasso (original approach), ElasticNet, Ridge, MI-based or tree-based methods like GENIE3). More confident TF-TG regulatory links will be inferred by more models and/or have higher performance scores (e.g. higher coefficient magnitudes $|c^*|$).

4.2.2.6 Incorporating more prior knowledge into the model

We can incorporate other regulatory interactions among TF proteins from various data resources besides PPI databases. That is, we can also consider integrating other important biological knowledge in our networks for regularization, such as signaling pathways. In doing so, we may uncover more systematic insights of direct and indirect TF interactions (i.e. coordination among TFs) on gene regulation. If there are other networks that capture relationships among TF predictors, we could consider adding those networks in with other hyperparameters. We may also explore additional ways to add more prior knowledge to the network regularized regression term, such as enforcing positive coefficients for known activator TFs and negative coefficients for known repressor TFs.

4.2.2.7 Characterizing direct versus indirect coordination links

NetREm may uncover cell-type-specific and/or disease-specific PPI subnetworks of and the network properties and protein properties of these transcriptional regulatory modules (TRMs) of direct and/or indirect TF-TF interactions can be analyzed to help understand PPI network-disease relationships, an open area of future research (Sevimoglu and Arga 2014). NetREm helps characterize the existing global PPI networks of known PPIs, by identifying cell-type-specific TF-TF PPI subnetworks that coordinate to co-regulate TGs. We may characterize known TF-TF links found in the input PPI as being indirect or direct, to provide more details on their mode of interaction with one another. To this end, we may look at physical PPIs from STRINGdb and other resources and determine which PPIs are direct; the remaining known PPIs may be considered indirect PPIs. This remains a big yet important area to research on. In addition, NetREm identifies novel TF-TF PPIs for follow-ups to investigate. This is essential as only a few actual PPIs have been identified and a majority are still unknown. Future work for NetREm may also include uncovering the true nature of novel TF-TF coordination links and whether the novel interactions are indirect interactions or potential direct PPI links that have not yet been discovered. One plan would include utilizing tools based on DeepMind's AlphaFold protein prediction model, such as AlphaFold2 (Singh and Vig 2017) or AlphaFold3 (Abramson et. al, 2024) that predicts protein-ligand interactions and protein-nucleic acid interactions; such tools could be applied to our novel TF-TF links to help predict the class of the TF-TF coordination link: indirect or direct.

4.2.2.8 Improving TF-TF coordination networks

We will need to develop approaches to reduce the number of False Positive (FP) TF-TF coordination links that are uncovered. In addition, there may be False Negative (FN) results due to potential missing links in the input PPI. Inherently, PPIs contain a significant proportion of FP and FN results. Various computational analysis techniques could be utilized to rank the reliability of PPI links (Sevimoglu and Arga 2014), mainly reducing the FPs. While NetREm utilizes a comprehensive input PPI that is derived from many different resources, there may be additional tools that could be used to detect FNs and add those

links (Karagoz and Arga 2013; Chua and Wong 2008). There are new data sources and technologies for inferring cooperativity among TFs, which could be used as input data to NetREm's PPI or for validation of the model results and potential reduction of FP results. Traditional methods like ChIP-seq for identifying bound TFs are limited in resolution and sensitivity. However, newer chromatin profiling techniques, including ChIP-exo (Rao et al. 2021; He et al. 2015), ORGANIC native-ChIP (Kasinathan et. al 2014), CUT&RUN (Skene and Henikoff 2017), offer base-pair level detail. Additionally, DNase and micrococcal nuclease (MNase) methods help map in vivo TF footprints. DNase digestion mainly reveals accessibility, while MNase limit digest produces DNA fragments shielded by chromatin proteins (Hesselberth et. al 2009; Henikoff et al. 2011), aiding in genome-wide inference of accessibility and binding. There are new studies that integrate high-resolution MNase-seq, ORGANIC ChIP, CUT&RUN, and dSMF (dual-enzyme single-molecule footprinting) to pinpoint TF-binding events at enhancers in the S2 cells of Drosophila (fruit fly organism). A recent study devised a technique combining MNase-seq and CUT&RUN to simultaneously map multiple TF bindings, deducing co-binding events for Drosophila (Rao et al. 2021). The unbound state of an enhancer is gauged using dSMF, to facilitate assessment of cooperativity among co-binding TFs at enhancers; the authors verify that co-binding is linked to nucleosome occupancy and stability, aligning with theories that TF cooperativity facilitates nucleosome displacement at active enhancers. Notably, the low occupancy of TFBSs in the Drosophila genome suggests that transient TF binding and slow nucleosome replacement are key to enhancer functionality. In the future, these tools may be applied to uncover cooperativity among TFs genome-wide in humans.

4.2.2.6 Additional benchmarking of NetREm's TF-TG regulatory networks

We benchmarked the performance of our complementary GRNs (our TF-TG regulatory links) with that of other well-known, state-of-the-art methods. For Schwann cells (SCs) we have validation data for 8 core TFs, which we explored. We gathered metrics on validated direct TGs, loss of function TGs, and novel candidate TGs for these 8 TFs in myelinating Schwann cells and non-myelinating SCs based on NetREm predictions. We used SERGIO to simulate single-cell data in human Embryonic Stem Cells based on

ground truth input GRN; we can use the recent GRouNdGAN (Zinati et al. 2024) method (improves upon SERGIO tool) to generate improved gene expression datasets for our evaluation of our method. In addition, we can utilize recently predicted cell-type GRNs by brainSCOPE (Emani et al. 2024) to evaluate our predicted networks (in terms of predicting TF-TG pairs that are experimentally validated or computationally identified) in glial and neuronal cell-types in our AD versus control application.

4.2.2.7 Use TF-TF coordination networks to uncover modules of genes with similar patterns of coregulation

Multiple TGs may be co-regulated by common TFs. Co-expression network analysis is a valuable method used to identify co-regulated genes and discover new disease genes and gene co-expression modules. Typically, genes with strongly correlated expression profiles often share common TFs binding to their promoters and cluster into the same modules. These modules represent biological functions and paths associated with disease phenotypes. To reveal modules of genes with similar patterns of coordination, we may perform hierarchical clustering of TG-specific *B* coordination scores for the cell-type across the cell-types and/or disease states. In doing this, we may uncover clusters of genes (e.g. gene modules) based on their TF-TF coordination scores. We may associate TF-TF links with various traits and context provided by the Contextual PPI database (Kotlyar et. al 2022). Then, we can determine module-trait correlations, similar to how WGCNA (Langfelder and Horvath 2008) determines phenotype-associated modules. Thus, this approach can work in tandem with single-cell WGCNA (scWGCNA), which determines gene co-expression modules based on expression dynamic patterns; in our case, NetREm-derived gene modules would be based on TF-TF coordination dynamic patterns, which may provide another layer of understanding.

4.2.2.8 Applications of Coordination Scores

Studies observe that TF-TF PPIs may play core roles in mediating the long-range enhancer regulation and have suggested that these TF PPIs can be used as input features to machine learning models to improve the predictive accuracy of long-range interactions among regulatory elements like enhancers and promoters (Wang et al. 2021a). Such models may prioritize TF PPIs for long-range enhancer regulation,

helping uncover properties of enhancer-driven gene regulation and biology. Nonetheless, the number of candidate TF PPIs is astronomical (~200,000), even after filtering for cell-type-specific TFs, resulting in high-dimensionality of TF PPI features and a strong risk of overfitting. NetREm enables discovery of cell-type-specific subnetworks of the PPIs, i.e. cell-type TF PPIs, and coordination scores B, which could help future downstream analysis work (e.g. predicting these long-range interactions among regulatory elements, especially at the cell-type-level, given the unique epigenomic landscape specific to cell-types).

4.2.2.9 Web-application for NetREm

We plan to convert NetREm findings to a web-application and web database for the community to use. In addition, we have deployed our open-source code on GitHub and will be providing more tutorials. We may add a standalone web-application to help users run NetREm seamlessly using their own input data.

§ 4.3 Concluding statements

I vividly recall a high school dinner conversation with my aunt, a Ph.D. in a biology-related field. She spread her arms wide and said, "If the true biological knowledge about a human cell is this vast, then we only know this tiny amount," indicating just a small inch. This realization, along with noticing in my biology textbook that the roles of centrioles were still uncertain, left me both fascinated and awed by how much remains unknown in this field. I recognize that my research is built upon the work of countless researchers who have paved the way with their dedication to expanding our understanding.

As I conclude this chapter of my academic journey, I am deeply committed to carrying forward the research and knowledge I've gained to make a meaningful impact in public health. My work reflects a lifelong mission to use science and data to improve lives and contribute to the betterment of society. With unwavering passion, empathy, and dedication, I look forward to the opportunities ahead to advance our understanding of diseases and contribute to the broader scientific community. This research is just the beginning, and I am eager to apply the lessons, experiences, and tools that I developed during this journey

(NetREm and SNPheno) to the next chapter of my career, to help make a lasting impact in the world. 😊

🕉 Om!

References

Abbasi J. 2022. Even Mild COVID-19 May Change the Brain. *JAMA* **327**: 1321–1322.

Abramson et. al, 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3 | Nature. https://www.nature.com/articles/s41586-024-07487-w (Accessed May 27, 2024).

Acquaah-Mensah GK, Agu N, Khan T, Gardner A. 2015. A Regulatory Role for the Insulin- and BDNF-Linked RORA in the Hippocampus: Implications for Alzheimer's Disease. *Journal of Alzheimer's Disease* **44**: 827–838.

Ahsendorf T, Müller F-J, Topkar V, Gunawardena J, Eils R. 2017. Transcription factors, coregulators, and epigenetic marks are linearly correlated and highly redundant. *PLOS ONE* **12**: e0186324.

Aibar et. al 2017. SCENIC: single-cell regulatory network inference and clustering | Nature Methods. https://www.nature.com/articles/nmeth.4463 (Accessed March 27, 2023).

Aiello Bowles EJ, Crane PK, Walker RL, Chubak J, LaCroix AZ, Anderson ML, Rosenberg D, Keene CD, Larson EB. 2019. Cognitive Resilience to Alzheimer's Disease Pathology in the Human Brain. *J Alzheimers Dis* **68**: 1071–1083.

Allocco DJ, Kohane IS, Butte AJ. 2004. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* **5**: 18.

Alzheimers.net. Alzheimer's Statistics | Alzheimers.net. https://www.alzheimers.net/alzheimers-statistics (Accessed November 7, 2023).

AMP-AD (Agora). Gene Comparison | Visual comparison tool for AD genes. https://agora.adknowledgeportal.org/genes/comparison (Accessed November 7, 2023).

Amruta et. al 2021. SARS-CoV-2 mediated neuroinflammation and the impact of COVID-19 in neurological disorders - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1359610121000186?via%3Dihub (Accessed November 7, 2023).

Amruta N, Chastain WH, Paz M, Solch RJ, Murray-Brown IC, Befeler JB, Gressett TE, Longo MT, Engler-Chiurazzi EB, Bix G. 2021. SARS-CoV-2 mediated neuroinflammation and the impact of COVID-19 in neurological disorders. *Cytokine Growth Factor Rev* **58**: 1–15.

Anderson et. al 2021. COVID-19 Tied to Acceleration of Alzheimer's Pathology. https://www.medscape.com/viewarticle/955755 (Accessed November 7, 2023).

Arment S SA <seth ament at, Shannon P PS <pshannon at, Richards M MR <mrichard at. 2021. trena: Fit transcriptional regulatory networks using gene expression, priors, machine learning. https://bioconductor.org/packages/trena/ (Accessed November 7, 2023).

AutismSpeaks.org. About Autism Speaks | Autism Speaks. https://www.autismspeaks.org/about-autism-speaks (Accessed April 18, 2024).

Avraham O, Chamessian A, Feng R, Yang L, Halevi AE, Moore AM, Gereau RWI, Cavalli V. 2022. Profiling the molecular signature of satellite glial cells at the single cell level reveals high similarities between rodents and humans. *PAIN* **163**: 2348.

Bader et. al, 2021. Home - Bader Lab @ The University of Toronto. https://baderlab.org/ (Accessed August 26, 2023).

Badia-i-Mompel P, Wessels L, Müller-Dott S, Trimbour R, Ramirez Flores RO, Argelaguet R, Saez-Rodriguez J. 2023. Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* **24**: 739–754.

Bakken et. al 2018. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types | PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209648 (Accessed November 14, 2023).

Balakrishnan A, Belfiore L, Chu T-H, Fleming T, Midha R, Biernaskie J, Schuurmans C. 2021. Insights Into the Role and Potential of Schwann Cells for Peripheral Nerve Repair From Studies of Development and Injury. *Front Mol Neurosci* **13**: 608442.

Bartsch et. al, 2011. CA1 neurons in the human hippocampus are critical for autobiographical memory, mental time travel, and autonoetic consciousness | PNAS. https://www.pnas.org/doi/10.1073/pnas.1110266108?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed (Accessed November 7, 2023).

Bellenguez et. al 2022. New insights into the genetic etiology of Alzheimer's disease and related dementias | Nature Genetics. https://www.nature.com/articles/s41588-022-01024-z (Accessed November 7, 2023).

Benedetti V, Banfi F, Zaghi M, Moll-Diaz R, Massimino L, Argelich L, Bellini E, Bido S, Muggeo S, Ordazzo G, et al. 2022. A SOX2-engineered epigenetic silencer factor represses the glioblastoma genetic program and restrains tumor development. *Science Advances* **8**: eabn3986.

Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. 2018. Religious Orders Study and Rush Memory and Aging Project. *J Alzheimers Dis* **64**: S161–S189.

Bentsen et. al 2022. TF-COMB – Discovering grammar of transcription factor binding sites - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2001037022003051 (Accessed March 27, 2023).

Berenson A, Lane R, Soto-Ugaldi LF, Patel M, Ciausu C, Li Z, Chen Y, Shah S, Santoso C, Liu X, et al. 2023. Paired yeast one-hybrid assays to detect DNA-binding cooperativity and antagonism across transcription factors. *Nat Commun* **14**: 6570.

Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, Vitale L, Pelleri MC, Tassani S, Piva F, et al. 2013. An estimation of the number of cells in the human body. *Annals of Human Biology* **40**: 463–471.

Blalock EM, Buechel HM, Popovic J, Geddes JW, Landfield PW. 2011. Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient Alzheimer's disease. *Journal of Chemical Neuroanatomy* **42**: 118–126.

Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW. 2004. Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences* **101**: 2173–2178.

Bossaerts L, Cacace R, Van Broeckhoven C. 2022. The role of ATP-binding cassette subfamily A in the etiology of Alzheimer's disease. *Molecular Neurodegeneration* **17**: 31.

Botía et. al 2017. An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks | BMC Systems Biology | Full Text. https://bmcsystbiol.biomedcentral.com/articles/10.1186/s12918-017-0420-6 (Accessed November 7, 2023).

Bowen EFW, Burgess JL, Granger R, Kleinman JE, Rhodes CH. 2019. DLPFC transcriptome defines two molecular subtypes of schizophrenia. *Transl Psychiatry* **9**: 1–10.

Boyle P, Després C. 2010. Dual-function transcription factors and their entourage. *Plant Signal Behav* **5**: 629–634.

Bravo González-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**: 1355–1367.

Brinton RD, Gore AC, Schmidt PJ, Morrison JH. 2009. 68 - Reproductive Aging of Females: Neural Systems. In *Hormones, Brain and Behavior (Second Edition)* (eds. D.W. Pfaff, A.P. Arnold, A.M. Etgen, S.E. Fahrbach, and R.T. Rubin), pp. 2199–2224, Academic Press, San Diego https://www.sciencedirect.com/science/article/pii/B9780080887838000681 (Accessed May 24, 2024).

Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, Aryee MJ, Majeti R, Chang HY, Greenleaf WJ. 2018. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**: 1535-1548.e16.

Byrois et. al 2022. Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders | Nature Neuroscience. https://www.nature.com/articles/s41593-022-01128-z (Accessed August 22, 2023).

Campos LM de, Cano A, Castellano JG, Moral S. 2019. Combining gene expression data and prior knowledge for inferring gene regulatory networks via Bayesian networks using structural restrictions. *Statistical Applications in Genetics and Molecular Biology* **18**. https://www.degruyter.com/document/doi/10.1515/sagmb-2018-0042/html (Accessed November 17, 2023).

Cappelletti C, Henriksen SP, Geut H, Rozemuller AJM, van de Berg WDJ, Pihlstrøm L, Toft M. 2023. Transcriptomic profiling of Parkinson's disease brains reveals disease stage specific gene expression changes. *Acta Neuropathol* **146**: 227–244.

Carlson et. al 2016. hgu133a.db. *Bioconductor*. http://bioconductor.org/packages/hgu133a.db/ (Accessed November 7, 2023).

Carpanini et. al 2019. Frontiers | Therapeutic Inhibition of the Complement System in Diseases of the Central Nervous System. https://www.frontiersin.org/articles/10.3389/fimmu.2019.00362/full (Accessed November 7, 2023).

Chagas VS, Groeneveld CS, Oliveira KG, Trefflich S, de Almeida RC, Ponder BAJ, Meyer KB, Jones SJM, Robertson AG, Castro MAA. 2019. RTNduals: an R/Bioconductor package for analysis of co-regulation and inference of dual regulons. *Bioinformatics* **35**: 5357–5358.

Chan TE, Stumpf MPH, Babtie AC. 2017. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *cels* **5**: 251-267.e3.

Chandrashekar et. al. 2023. DeepGAMI: deep biologically guided auxiliary learning for multimodal integration and imputation to improve genotype–phenotype prediction. https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-023-01248-6 (Accessed July 6, 2024).

Cheng F, Zhao J, Wang Y, Lu W, Liu Z, Zhou Y, Martin WR, Wang R, Huang J, Hao T, et al. 2021. Comprehensive characterization of protein-protein interactions perturbed by disease mutations. *Nat Genet* **53**: 342–353.

Cho et. al 2017. VEGFR2 alteration in Alzheimer's disease | Scientific Reports. https://www.nature.com/articles/s41598-017-18042-1 (Accessed November 7, 2023).

Choi E, Zhang X, Xing C, Yu H. 2016. Mitotic Checkpoint Regulators Control Insulin Signaling and Metabolic Homeostasis. *Cell* **166**: 567.

Choy CT, Wong CH, Chan SL. 2019. Embedding of Genes Using Cancer Gene Expression Data: Biological Relevance and Potential Application on Biomarker Discovery. *Frontiers in Genetics* **9**. https://www.frontiersin.org/articles/10.3389/fgene.2018.00682 (Accessed January 2, 2024).

Chu X, Guan B, Dai L, Liu J, Li F, Shang J. 2023. Network embedding framework for driver gene discovery by combining functional and structural information. *BMC Genomics* **24**: 426.

Chua and Wong 2008. Increasing the reliability of protein interactomes - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1359644608001645?via%3Dihub (Accessed November 15, 2023).

Coetzee SG, Coetzee GA, Hazelett DJ. 2015. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**: 3847–3849.

Congdon E, Poldrack RA, Freimer NB. 2010. Neurocognitive Phenotypes and Genetic Dissection of Disorders of Brain and Behavior. *Neuron* **68**: 218–230.

Corces MR, Shcherbina A, Kundu S, Gloudemans MJ, Frésard L, Granja JM, Louie BH, Eulalio T, Shams S, Bagdatli ST, et al. 2020. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat Genet* **52**: 1158–1168.

COVID-19 Host Genetics Initiative 2022. COVID-19 Host Genetics Initiative. https://www.covid19hg.org/results/r7/ (Accessed November 7, 2023).

Coyle PK. 2011. Dissecting the Immune Component of Neurologic Disorders: A Grand Challenge for the 21st Century. *Front Neurol* **2**: 37.

Cuomo ASE, Nathan A, Raychaudhuri S, MacArthur DG, Powell JE. 2023. Single-cell genomics meets human genetics. *Nat Rev Genet* **24**: 535–549.

D'Alessio AC, Fan ZP, Wert KJ, Baranov P, Cohen MA, Saini JS, Cohick E, Charniga C, Dadon D, Hannett NM, et al. 2015. A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity. *Stem Cell Reports* **5**: 763–775.

Davis et. al 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/23/14/1846/190290?login=false (Accessed November 7, 2023).

De Las Rivas J, Fontanillo C. 2012. Protein–protein interaction networks: unraveling the wiring of molecular machines within the cell. *Briefings in Functional Genomics* **11**: 489–496.

del Re EC, Spencer KM, Oribe N, Mesholam-Gately RI, Goldstein J, Shenton ME, Petryshen T, Seidman LJ, McCarley RW, Niznikiewicz MA. 2015. Clinical high risk and first episode schizophrenia: Auditory event-related potentials. *Psychiatry Res* **231**: 126–133.

Diaz Brinton R. 2012. Minireview: Translational Animal Models of Human Menopause: Challenges and Emerging Opportunities. *Endocrinology* **153**: 3571–3578.

Dibaeinia P, Sinha S. 2020. SERGIO: A Single-Cell Expression Simulator Guided by Gene Regulatory Networks. *cels* **11**: 252-271.e11.

Ding R, Wang Q, Gong L, Zhang T, Zou X, Xiong K, Liao Q, Plass M, Li L. 2023. scQTLbase: an integrated human single-cell eQTL database. *Nucleic Acids Research* gkad781.

Diogo D, Tian C, Franklin CS, Alanne-Kinnunen M, March M, Spencer CCA, Vangjeli C, Weale ME, Mattsson H, Kilpeläinen E, et al. 2018. Phenome-wide association studies across large population cohorts support drug target validation. *Nat Commun* **9**: 4285.

Dirmeier S, Fuchs C, Mueller NS, Theis FJ. 2018. netReg: network-regularized linear models for biological association studies. *Bioinformatics* **34**: 896–898.

Dubey J, Ratnakaran N, Koushika SP. 2015. Neurodegeneration and microtubule dynamics: death by a thousand cuts. *Front Cell Neurosci* **9**: 343.

Duttke SH, Guzman C, Chang M, Delos Santos NP, McDonald BR, Xie J, Carlin AF, Heinz S, Benner C. 2024. Position-dependent function of human sequence-specific transcription factors. *Nature* **631**: 891–898.

Ebanks B, Ingram TL, Chakrabarti L. 2020. ATP synthase and Alzheimer's disease: putting a spin on the mitochondrial hypothesis. *Aging (Albany NY)* **12**: 16647–16662.

Emani PS, Liu JJ, Clarke D, Jensen M, Warrell J, Gupta C, Meng R, Lee CY, Xu S, Dursun C, et al. 2024. Single-cell genomics and regulatory networks for 388 human brains. *Science* **384**: eadi5199.

Eraslan et. al 2022. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function | Science. https://www.science.org/doi/10.1126/science.abl4290 (Accessed April 5, 2024).

Erausquin et. al 2021. The chronic neuropsychiatric sequelae of COVID-19: The need for a prospective study of viral impact on brain functioning - Erausquin - 2021 - Alzheimer's & Dementia - Wiley Online Library. https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.12255 (Accessed November 7, 2023).

Escorcia-Rodríguez JM, Gaytan-Nuñez E, Hernandez-Benitez EM, Zorro-Aranda A, Tello-Palencia MA, Freyre-González JA. 2023. Improving gene regulatory network inference and assessment: The importance of using network structure. *Frontiers in Genetics* **14**. https://www.frontiersin.org/articles/10.3389/fgene.2023.1143382 (Accessed November 24, 2023).

Eyigoz E, Mathur S, Santamaria M, Cecchi G, Naylor M. 2020. Linguistic markers predict onset of Alzheimer's disease. *eClinicalMedicine* **28**. https://www.thelancet.com/journals/eclinm/article/PIIS2589-53702030327-8/fulltext (Accessed September 6, 2023).

Fagone et. al 2020. Transcriptional landscape of SARS-CoV-2 infection dismantles pathogenic pathways activated by the virus, proposes unique sex-specific differences and predicts tailored therapeutic strategies - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1568997220301336?via%3Dihub (Accessed November 7, 2023).

Fahira A, Li Z, Liu N, Shi Y. 2019. Prediction of causal genes and gene expression analysis of attention-deficit hyperactivity disorder in the different brain region, a comprehensive integrative analysis of ADHD. *Behavioural Brain Research* **364**: 183–192.

Fan et. al 2013. RMA normalization for microarray data. https://felixfan.github.io/RMA-Normalization-Microarray/ (Accessed November 7, 2023).

Fang L-P, Zhao N, Caudal LC, Chang H-F, Zhao R, Lin C-H, Hainz N, Meier C, Bettler B, Huang W, et al. 2022. Impaired bidirectional communication between interneurons and oligodendrocyte precursor cells affects social cognitive behavior. *Nat Commun* **13**: 1394.

Farrow SL, Schierding W, Gokuladhas S, Golovina E, Fadason T, Cooper AA, O'Sullivan JM. 2022. Establishing gene regulatory networks from Parkinson's disease risk loci. *Brain* **145**: 2422–2435.

Federico A, Monti S. 2020. Contextualized Protein-Protein Interactions. *Patterns (N Y)* **2**: 100153.

Fröb F, Wegner M. 2021. Coordination of Schwann cell myelination and node formation at the transcriptional level. *Neural Regen Res* **17**: 1269–1270.

Fullard et. al 2018. An atlas of chromatin accessibility in the adult human brain. https://genome.cshlp.org/content/28/8/1243.long (Accessed November 7, 2023).

Gabbouj et. al 2019. Frontiers | Altered Insulin Signaling in Alzheimer's Disease Brain – Special Emphasis on PI3K-Akt Pathway. https://www.frontiersin.org/articles/10.3389/fnins.2019.00629/full (Accessed November 7, 2023).

Gandal et. al 2018. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder | Science. https://www.science.org/doi/10.1126/science.aat8127 (Accessed April 21, 2024).

Gautier et. al, 2004. affy—analysis of Affymetrix GeneChip data at the probe level | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/20/3/307/185980?login=true (Accessed November 7, 2023).

GBD 2016 Neurology Collaborators et. al 2019. Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016 - The Lancet Neurology. https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(18)30499-X/fulltext (Accessed November 10, 2023).

Gerber D, Pereira JA, Gerber J, Tan G, Dimitrieva S, Yángüez E, Suter U. 2021. Transcriptional profiling of mouse peripheral nerves to the single-cell level to build a sciatic nerve ATlas (SNAT) eds. D.E. Bergles, M.E. Bronner, L.A. Goff, and A. Hoke. *eLife* **10**: e58591.

Ghanbari M, Lasserre J, Vingron M. 2015. Reconstruction of gene networks using prior knowledge. *BMC Systems Biology* **9**: 84.

Gharavi E, Gu A, Zheng G, Smith JP, Cho HJ, Zhang A, Brown DE, Sheffield NC. 2021. Embeddings of genomic region sets capture rich biological associations in lower dimensions. *Bioinformatics* **37**: 4299–4306.

Glasgow SM, Zhu W, Stolt CC, Huang T-W, Chen F, LoTurco JJ, Neul JL, Wegner M, Mohila C, Deneen B. 2014. Mutual antagonism between Sox10 and NFIA regulates diversification of glial lineages and glioma subtypes. *Nat Neurosci* **17**: 1322–1329.

Goldstein IS, Erickson DJ, Sleeper LA, Haynes RL, Kinney HC. 2017. The Lateral Temporal Lobe in Early Human Life. *Journal of Neuropathology & Experimental Neurology* **76**: 424–438.

Gonzalez-Teran B, Pittman M, Felix F, Thomas R, Richmond-Buccola D, Hüttenhain R, Choudhary K, Moroni E, Costa MW, Huang Y, et al. 2022. Transcription factor protein interactomes reveal genetic determinants in heart disease. *Cell* **185**: 794-814.e30.

Göös H, Kinnunen M, Salokas K, Tan Z, Liu X, Yadav L, Zhang Q, Wei G-H, Varjosalo M. 2022. Human transcription factor protein interaction networks. *Nat Commun* **13**: 766.

Gordân R, Hartemink AJ, Bulyk ML. 2009. Distinguishing direct versus indirect transcription factor–DNA interactions. *Genome Res* **19**: 2090–2100.

Gordon et. al 2021. Impact of COVID-19 on the Onset and Progression of Alzheimer's Disease and Related Dementias: A Roadmap for Future Research - Gordon - 2022 - Alzheimer's & Dementia - Wiley Online Library. https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.12488 (Accessed November 7, 2023).

Goyal P, Ferrara E. 2018. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems* **151**: 78–94.

Greenfield et. al 2013. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/29/8/1060/232957 (Accessed August 23, 2024).

Groeneveld C, Robertson G, Wang X, Fletcher M, Markowetz F, Meyer K, Castro M. 2023. RTN: RTN: Reconstruction of Transcriptional regulatory Networks and analysis of regulons. https://bioconductor.org/packages/RTN/ (Accessed November 7, 2023).

Grote S, Prüfer K, Kelso J, Dannemann M. 2016. ABAEnrichment: an R package to test for gene set expression enrichment in the adult and developing human brain. *Bioinformatics* **32**: 3201–3203.

Grover A, Leskovec J. 2016. node2vec: Scalable Feature Learning for Networks. http://arxiv.org/abs/1607.00653 (Accessed September 3, 2023).

Gu et. al 2022. GEO Accession viewer. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM6466265 (Accessed April 11, 2024).

Gu Q, Nagaraj SH, Hudson NJ, Dalrymple BP, Reverter A. 2011. Genome-wide patterns of promoter sharing and co-expression in bovine skeletal muscle. *BMC Genomics* **12**: 23.

Guo et. al, 2019. Evaluation of Peripheral Immune Dysregulation in Alzheimer's Disease and Vascular Dementia - IOS Press. https://content.iospress.com/articles/journal-of-alzheimers-disease/jad190666 (Accessed November 7, 2023).

Guo J, Cai Y, Ye X, Ma N, Wang Y, Yu B, Wan J. 2019. MiR-409-5p as a Regulator of Neurite Growth Is Down Regulated in APP/PS1 Murine Model of Alzheimer's Disease. *Frontiers in Neuroscience* **13**. https://www.frontiersin.org/articles/10.3389/fnins.2019.01264 (Accessed November 7, 2023).

Guo Y, Gifford DK. 2017. Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. *BMC Genomics* **18**: 45.

Gupta C, Xu J, Jin T, Khullar S, Liu X, Alatkar S, Cheng F, Wang D. 2022. Single-cell network biology characterizes cell type gene regulation for drug repurposing and phenotype prediction in Alzheimer's disease. *PLOS Computational Biology* **18**: e1010287.

Gupta et. al 2022. Bringing machine learning to research on intellectual and developmental disabilities: taking inspiration from neurological diseases | Journal of Neurodevelopmental Disorders | Full Text. https://jneurodevdisorders.biomedcentral.com/articles/10.1186/s11689-022-09438-w (Accessed July 19, 2024).

Gutierrez-Sacristan A, Hern C, ez-Ferrer, Gonzalez JR, Furlong LI. 2023. psygenet2r: psygenet2r - An R package for querying PsyGeNET and to perform comorbidity studies in psychiatric disorders. https://bioconductor.org/packages/psygenet2r/ (Accessed November 7, 2023).

Halks-Miller M, Schroeder ML, Haroutunian V, Moenning U, Rossi M, Achim C, Purohit D, Mahmoudi M, Horuk R. 2003. CCR1 is an early and specific marker of Alzheimer's disease. https://onlinelibrary.wiley.com/doi/10.1002/ana.10733 (Accessed August 23, 2024).

Hampel et. al 2020. Frontiers | A Path Toward Precision Medicine for Neuroinflammatory Mechanisms in Alzheimer's Disease. https://www.frontiersin.org/articles/10.3389/fimmu.2020.00456/full (Accessed November 7, 2023).

Hannenhalli S, Levy S. 2002. Predicting transcription factor synergism. *Nucleic Acids Res* **30**: 4278–4284.

Haury A-C, Mordelet F, Vera-Licona P, Vert J-P. 2012. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Systems Biology* **6**: 145.

He C, Kalafut NC, Sandoval SO, Risgaard R, Sirois CL, Yang C, Khullar S, Suzuki M, Huang X, Chang Q, et al. 2023. BOMA, a machine-learning framework for comparative gene expression analysis across brains and organoids. *Cell Reports Methods* **3**. https://www.cell.com/cell-reports-methods/abstract/S2667-2375(23)00020-6 (Accessed July 19, 2024).

He Q, Johnston J, Zeitlinger J. 2015. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat Biotechnol* **33**: 395–401.

Heffron et. al, 2021. The landscape of antibody binding in SARS-CoV-2 infection | PLOS Biology. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001265 (Accessed November 7, 2023).

Heming M, Li X, Räuber S, Mausberg AK, Börsch A-L, Hartlehnert M, Singhal A, Lu I-N, Fleischer M, Szepanowski F, et al. 2021. Neurological Manifestations of COVID-19 Feature T Cell Exhaustion and Dedifferentiated Monocytes in Cerebrospinal Fluid. *Immunity* **54**: 164-175.e6.

Hemonnot et. al 2019. Frontiers | Microglia in Alzheimer Disease: Well-Known Targets and New Opportunities. https://www.frontiersin.org/articles/10.3389/fnagi.2019.00233/full (Accessed November 7, 2023).

Heneka MT, Carson MJ, El Khoury J, Landreth GE, Brosseron F, Feinstein DL, Jacobs AH, Wyss-Coray T, Vitorica J, Ransohoff RM, et al. 2015. Neuroinflammation in Alzheimer's Disease. *Lancet Neurol* **14**: 388–405.

Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S. 2011. Epigenome characterization at single base-pair resolution. *Proceedings of the National Academy of Sciences* **108**: 18318–18323.

Hesselberth et. al 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting | Nature Methods. https://www.nature.com/articles/nmeth.1313 (Accessed November 14, 2023).

Hoefsloot HCJ, Smit S, Smilde AK. 2008. A Classification Model for the Leiden Proteomics Competition. *Statistical Applications in Genetics and Molecular Biology* **7**. https://www.degruyter.com/document/doi/10.2202/1544-6115.1351/html (Accessed November 17, 2023).

Hou Y, Zhao J, Martin W, Kallianpur A, Chung MK, Jehi L, Sharifi N, Erzurum S, Eng C, Cheng F. 2020. New insights into genetic susceptibility of COVID-19: an ACE2 and TMPRSS2 polymorphism analysis. *BMC Medicine* **18**: 216.

Hsu Y-HH, Nacu E, Liu R, Pintacuda G, Kim A, Tsafou K, Petrossian N, Crotty W, Suh JM, Riseman J, et al. 2022. Using brain cell-type-specific protein interactomes to interpret genetic data in schizophrenia. 2021.10.07.21264568. https://www.medrxiv.org/content/10.1101/2021.10.07.21264568v2 (Accessed August 22, 2023).

Hu et. al, 2021. Genetic variants are identified to increase risk of COVID-19 related mortality from UK Biobank data | Human Genomics | Full Text. https://humgenomics.biomedcentral.com/articles/10.1186/s40246-021-00306-7 (Accessed November 7, 2023).

Hu S, Metcalf E, Mahat DB, Chan L, Sohal N, Chakraborty M, Hamilton M, Singh A, Singh A, Lees JA, et al. 2022. Transcription factor antagonism regulates heterogeneity in embryonic stem cell states. *Molecular Cell* **82**: 4410-4427.e12.

Hung HA, Sun G, Keles S, Svaren J. 2015. Dynamic Regulation of Schwann Cell Enhancers after Peripheral Nerve Injury *. *Journal of Biological Chemistry* **290**: 6937–6950.

Huynh-Thu et. al, 2010. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods | PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0012776 (Accessed November 7, 2023).

Ibarra et. al 2020. Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions | Nature Communications. https://www.nature.com/articles/s41467-019-13888-7 (Accessed March 27, 2023).

Iglesias-Martinez LF, Kolch W, Santra T. 2016. BGRMI: A method for inferring gene regulatory networks from time-course gene expression data and its application in breast cancer research. *Sci Rep* **6**: 37140.

Imoto S, Higuchi T, Goto T, Tashiro K, Kuhara S, Miyano S. 2003. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Proc IEEE Comput Soc Bioinform Conf* **2**: 104–113.

Inal J. 2020. Biological Factors Linking ApoE ε4 Variant and Severe COVID-19. *Curr Atheroscler Rep* **22**: 70.

Inestrosa and Varela-Nallar 2014. Wnt signaling in the nervous system and in Alzheimer's disease | Journal of Molecular Cell Biology | Oxford Academic. https://academic.oup.com/jmcb/article/6/1/64/874321?login=true (Accessed November 7, 2023).

Ishii A, Furusho M, Bansal R. 2021. Mek/ERK1/2-MAPK and PI3K/Akt/mTOR signaling plays both independent and cooperative roles in Schwann cell differentiation, myelination and dysmyelination. *Glia* **69**: 2429–2446.

Ivleva EI, Morris DW, Moates AF, Suppes T, Thaker GK, Tamminga CA. 2010. Genetics and intermediate phenotypes of the schizophrenia—bipolar disorder boundary. *Neuroscience & Biobehavioral Reviews* **34**: 897–921.

Jain A, Tuteja G. 2019. TissueEnrich: Tissue-specific gene enrichment analysis. *Bioinformatics* **35**: 1966–1967.

Jalkanen J, Khan S, Elima K, Huttunen T, Wang N, Hollmén M, Elo LL, Jalkanen S. 2023. Polymorphism in interferon alpha/beta receptor contributes to glucocorticoid response and outcome of ARDS and COVID-19. *Critical Care* **27**: 112.

Jansen et. al 2019. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk | Nature Genetics. https://www.nature.com/articles/s41588-018-0311-9 (Accessed November 7, 2023).

Java et. al, 2020. JCI Insight - The complement system in COVID-19: friend and foe? https://insight.jci.org/articles/view/140711 (Accessed November 7, 2023).

Jha et. al 2019. Nuclear factor-kappa β as a therapeutic target for Alzheimer's disease - Jha - 2019 - Journal of Neurochemistry - Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1111/jnc.14687 (Accessed November 7, 2023).

Jha K, Saha S, Singh H. 2022. Prediction of protein–protein interaction using graph neural networks. *Sci Rep* **12**: 8360.

Jia L, Li F, Wei C, Zhu M, Qu Q, Qin W, Tang Y, Shen L, Wang Y, Shen L, et al. 2020. Prediction of Alzheimer's disease using multi-variants from a Chinese genome-wide association study. *Brain* **144**: 924–937.

Jiang et. al 2020. scREAD: A Single-Cell RNA-Seq Database for Alzheimer's Disease - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2589004220309664?via%3Dihub (Accessed November 7, 2023).

Jin T, Rehani P, Ying M, Huang J, Liu S, Roussos P, Wang D. 2021. scGRNom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. *Genome Medicine* **13**: 95.

Johnson KL, Qi Z, Yan Z, Wen X, Nguyen TC, Zaleta-Rivera K, Chen C-J, Fan X, Sriram K, Wan X, et al. 2021. Revealing protein-protein interactions at the transcriptome scale by sequencing. *Molecular Cell* **81**: 4091-4103.e9.

Jones EA, Jang S-W, Mager GM, Chang L-W, Srinivasan R, Gokey NG, Ward RM, Nagarajan R, Svaren J. 2007. Interactions of Sox10 and Egr2 in Myelin Gene Regulation. *Neuron Glia Biol* **3**: 377–387.

Joukov V, De Nicolo A, Rodriguez A, Walter JC, Livingston DM. 2010. Centrosomal protein of 192 kDa (Cep192) promotes centrosome-driven spindle assembly by engaging in organelle-specific Aurora A activation. *Proceedings of the National Academy of Sciences* **107**: 21022–21027.

Joung J, Ma S, Tay T, Geiger-Schuller KR, Kirchgatterer PC, Verdine VK, Guo B, Arias-Garcia MA, Allen WE, Singh A, et al. 2023. A transcription factor atlas of directed differentiation. *Cell* **186**: 209-229.e26.

Ju H, Yun H, Kim Y, Nam YJ, Lee S, Lee J, Jeong SM, Heo J, Kwon H, Cho YS, et al. 2023. Activating transcription factor-2 supports the antioxidant capacity and ability of human mesenchymal stem cells to prevent asthmatic airway inflammation. *Exp Mol Med* **55**: 413–425.

Jung et. al 2019. A compendium of promoter-centered long-range chromatin interactions in the human genome | Nature Genetics. https://www.nature.com/articles/s41588-019-0494-8 (Accessed November 7, 2023).

Kamimoto K, Stringa B, Hoffmann CM, Jindal K, Solnica-Krezel L, Morris SA. 2023. Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**: 742–751.

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**: 27–30.

Karagoz and Arga 2013. Assessment of high-confidence protein–protein interactome in yeast - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1476927113000212?via%3Dihub (Accessed November 15, 2023).

Karczewski KJ, Tatonetti NP, Landt SG, Yang X, Slifer T, Altman RB, Snyder M. 2011. Cooperative transcription factor associations discovered using regulatory variation. *Proc Natl Acad Sci USA* **108**: 13353–13358.

Kasinathan et. al 2014. High-resolution mapping of transcription factor binding sites on native chromatin | Nature Methods. https://www.nature.com/articles/nmeth.2766 (Accessed November 14, 2023).

Kellett and Hooper 2009. Full article: Prion protein and Alzheimer disease. https://www.tandfonline.com/doi/full/10.4161/pri.3.4.9980 (Accessed November 7, 2023).

Khullar R, Shah S, Singh G, Bae J, Gattu R, Jain S, Green J, Anandarangam T, Cohen M, Madan N, et al. 2020. Effects of Prone Ventilation on Oxygenation, Inflammation, and Lung Infiltrates in COVID-19 Related Acute Respiratory Distress Syndrome: A Retrospective Cohort Study. *Journal of Clinical Medicine* **9**: 4129.

Khullar S, Huang X, Ramesh R, Svaren J, Wang D. 2023. NetREm: Network Regression Embeddings reveal cell-type transcription factor coordination for gene regulation. 2023.10.25.563769. https://www.biorxiv.org/content/10.1101/2023.10.25.563769v1 (Accessed October 30, 2023).

Khullar S, Wang D. 2023. Predicting brain-regional gene regulatory networks from multi-omics for Alzheimer's disease phenotypes and Covid-19 severity. *Human Molecular Genetics* **32**: 1797–1813.

Khullar S, Wang D. 2021. Predicting gene regulatory networks from multi-omics to link genetic risk variants and neuroimmunology to Alzheimer's disease phenotypes. 2021.06.21.449165. https://www.biorxiv.org/content/10.1101/2021.06.21.449165v2 (Accessed July 19, 2024).

Kim D, Tran A, Kim HJ, Lin Y, Yang JYH, Yang P. 2023. Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data. *npj Syst Biol Appl* **9**: 1–13.

Kim S, Pan W, Shen X. 2013. Network-based penalized regression with application to genomic data. *Biometrics* **69**: 582–593.

Kinney JW, Bemiller SM, Murtishaw AS, Leisgang AM, Salazar AM, Lamb BT. 2018a. Inflammation as a central mechanism in Alzheimer's disease. *Alzheimers Dement (N Y)* **4**: 575–590.

Kinney JW, Bemiller SM, Murtishaw AS, Leisgang AM, Salazar AM, Lamb BT. 2018b. Inflammation as a central mechanism in Alzheimer's disease. *Alzheimers Dement (N Y)* **4**: 575–590.

Kircheis R, Haasbach E, Lueftenegger D, Heyken WT, Ocker M, Planz O. 2020. NF-κB Pathway as a Potential Target for Treatment of Critical Stage COVID-19 Patients. *Frontiers in Immunology* **11**. https://www.frontiersin.org/articles/10.3389/fimmu.2020.598444 (Accessed November 7, 2023).

Klemm et. al, 2019. Chromatin accessibility and the regulatory epigenome | Nature Reviews Genetics. https://www.nature.com/articles/s41576-018-0089-8 (Accessed November 28, 2023).

Kong Y, Han J, Wu X, Zeng H, Liu J, Zhang H. 2020. VEGF-D: a novel biomarker for detection of COVID-19 progression. *Critical Care* **24**: 373.

Kong Y, Yu T. 2018. A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data. *Bioinformatics* **34**: 3727–3737.

Kotlyar et. al 2022. IID 2021: towards context-specific protein interaction analyses by increased coverage, enhanced annotation and enrichment analysis | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/50/D1/D640/6424757 (Accessed July 14, 2023).

Kumar et. al 2017. Extent of Dorsolateral Prefrontal Cortex Plasticity and Its Association With Working Memory in Patients With Alzheimer Disease | Dementia and Cognitive Impairment | JAMA Psychiatry | JAMA Network. https://jamanetwork.com/journals/jamapsychiatry/fullarticle/2658231 (Accessed November 7, 2023).

Kumar et. al, 2020. Dorsolateral prefrontal cortex metabolites and their relationship with plasticity in Alzheimer's disease - Kumar - 2020 - Alzheimer's & Dementia - Wiley Online Library. https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.045879 (Accessed May 24, 2024).

Kumar S, Ambrosini G, Bucher P. 2017. SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res* **45**: D139–D144.

Kunkle et. al 2019. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing | Nature Genetics. https://www.nature.com/articles/s41588-019-0358-2 (Accessed November 7, 2023).

Kwee TC, Kwee RM. 2020. Chest CT in COVID-19: What the Radiologist Needs to Know. *RadioGraphics* **40**: 1848–1865.

Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, Duong TE, Gao D, Chun J, Kharchenko PV, et al. 2018. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**: 70–80.

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. *Cell* **172**: 650–665.

Landhuis E. 2021. Could the immune system be key to Alzheimer's disease? *Knowable Magazine | Annual Reviews*. https://knowablemagazine.org/article/health-disease/2021/could-immune-system-be-key-alzheimers-disease (Accessed November 7, 2023).

Langfelder and Horvath 2008. WGCNA: an R package for weighted correlation network analysis | BMC Bioinformatics | Full Text. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559 (Accessed November 7, 2023).

Lawrence T. 2009. The Nuclear Factor NF-κB Pathway in Inflammation. *Cold Spring Harb Perspect Biol* **1**: a001651.

Lear BP, Thompson EAN, Rodriguez K, Arndt ZP, Khullar S, Klosa PC, Lu RJ, Morrow CS, Risgaard R, Peterson ER, et al. 2023. Age-maintained human neurons demonstrate a developmental loss of intrinsic neurite growth ability. 2023.05.23.541995. https://www.biorxiv.org/content/10.1101/2023.05.23.541995v1 (Accessed July 6, 2024).

LeBlanc et. al 2006. Full article: Neuropathy-Associated Egr2 Mutants Disrupt Cooperative Activation of Myelin Protein Zero by Egr2 and Sox10. https://www.tandfonline.com/doi/full/10.1128/MCB.01689-06 (Accessed November 2, 2023).

Lee and Kim 2017. Molecules | Free Full-Text | Recent Advances in the Inhibition of p38 MAPK as a Potential Strategy for the Treatment of Alzheimer's Disease. https://www.mdpi.com/1420-3049/22/8/1287 (Accessed November 7, 2023).

Lee B-K, Bhinge AA, Battenhouse A, McDaniell RM, Liu Z, Song L, Ni Y, Birney E, Lieb JD, Furey TS, et al. 2012. Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome Res* **22**: 9–24.

Leng F, Edison P. 2021. Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here? *Nat Rev Neurol* **17**: 157–172.

Li and Li 2008. Network-constrained regularization and variable selection for analysis of genomic data | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/24/9/1175/206444 (Accessed March 27, 2023).

Li B, Ritchie MD. 2021. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Front Genet* **12**: 713230.

Li C, Li H. 2010. VARIABLE SELECTION AND REGRESSION ANALYSIS FOR GRAPH-STRUCTURED COVARIATES WITH AN APPLICATION TO GENOMICS. *Ann Appl Stat* **4**: 1498–1516.

Li L, Liu Z-P. 2022. A connected network-regularized logistic regression model for feature selection. *Appl Intell* **52**: 11672–11702.

Li Y, Jackson SA. 2015. Gene Network Reconstruction by Integration of Prior Biological Knowledge. *G3 (Bethesda)* **5**: 1075–1079.

Li Y, Xiao X, Chen H, Chen Z, Hu K, Yin D. 2020. Transcription factor NFYA promotes G1/S cell cycle transition and cell proliferation by transactivating cyclin D1 and CDK4 in clear cell renal cell carcinoma. *Am J Cancer Res* **10**: 2446–2463.

Lindsey et. al 2022. Association of COVID-19 with New-Onset Alzheimer's Disease - IOS Press. https://content.iospress.com/articles/journal-of-alzheimers-disease/jad220717 (Accessed November 7, 2023).

Liu et. al 2014. Neuroinflammation in Alzheimer's disease: chemokines produced by astrocytes and chemokine receptors - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4314046/ (Accessed August 23, 2024).

Liu et. al 2022. Illuminating links between cis-regulators and trans-acting variants in the human prefrontal cortex | Genome Medicine | Full Text. https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-022-01133-8 (Accessed May 27, 2024).

Liu M, Guo S, Hibbert JM, Jain V, Singh N, Wilson NO, Stiles JK. 2011. CXCL10/IP-10 in infectious diseases pathogenesis and potential therapeutic implications. *Cytokine & Growth Factor Reviews* **22**: 121–130.

Liu N, Xu J, Liu H, Zhang S, Li M, Zhou Y, Qin W, Li MJ, Yu C, Initiative for the A disease N. 2021. Hippocampal transcriptome-wide association study and neurobiological pathway analysis for Alzheimer's disease. *PLOS Genetics* **17**: e1009363. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009363 (Accessed June 12, 2021).

Liu X, Salokas K, Tamene F, Jiu Y, Weldatsadik RG, Öhman T, Varjosalo M. 2018. An AP-MS- and BioID-compatible MAC-tag enables comprehensive mapping of protein interactions and subcellular localizations. *Nat Commun* **9**: 1188.

Lopez-Anido C, Poitelon Y, Gopinath C, Moran JJ, Ma KH, Law WD, Antonellis A, Feltri ML, Svaren J. 2016. Tead1 regulates the expression of Peripheral Myelin Protein 22 during Schwann cell development. *Human Molecular Genetics* **25**: 3055–3069.

Lopez-Anido et. al 2015. GEO Accession viewer. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64703 (Accessed July 14, 2023).

Love M, Ahlmann-Eltze C, Forbes K, Anders S, Huber W, FP7 RE, NHGRI N, CZI. 2023. DESeq2: Differential gene expression analysis based on the negative binomial distribution. https://bioconductor.org/packages/DESeq2/ (Accessed November 7, 2023).

Lu et. al 2014. REST and stress resistance in ageing and Alzheimer's disease | Nature. https://www.nature.com/articles/nature13163 (Accessed November 7, 2023).

Lu T, Aron L, Zullo J, Pan Y, Kim H, Chen Y, Yang T-H, Kim H-M, Drake D, Liu XS, et al. 2014. REST and stress resistance in ageing and Alzheimer's disease. *Nature* **507**: 448–454.

Luo J, Xu P, Cao P, Wan H, Lv X, Xu S, Wang G, Cook MN, Jones BC, Lu L, et al. 2018. Integrating Genetic and Gene Co-expression Analysis Identifies Gene Networks Involved in Alcohol and Stress Responses. *Frontiers in Molecular Neuroscience* **11**: 102.

Luo W, Brouwer C. 2013. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29**: 1830–1831.

Ma KH, Svaren J. 2018. Epigenetic Control of Schwann Cells. *Neuroscientist* **24**: 627–638.

Maccioni et. al 2018. Alzheimer´s Disease in the Perspective of Neuroimmunology - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6040210/ (Accessed November 7, 2023).

Machiela et. al. 2015. LDLink: Web-based applications to interrogate linkage disequilibrium in populations - NCI. https://dceg.cancer.gov/tools/analysis/ldlink (Accessed August 27, 2023).

MacQueen et. al 2005. The phenotypes of bipolar disorder: relevance for genetic investigations | Molecular Psychiatry. https://www.nature.com/articles/4001701 (Accessed April 18, 2024).

Maezawa et. al 2012. Microglial KCa3.1 Channels as a Potential Therapeutic Target for Alzheimer's Disease. https://www.hindawi.com/journals/ijad/2012/868972/ (Accessed November 7, 2023).

Maezawa I, Jenkins DP, Jin BE, Wulff H. 2012. Microglial KCa3.1 Channels as a Potential Therapeutic Target for Alzheimer's Disease. *Int J Alzheimers Dis* **2012**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3364551/ (Accessed April 19, 2021).

Maglorius Renkilaraj MRL, Baudouin L, Wells CM, Doulazmi M, Wehrlé R, Cannaya V, Bachelin C, Barnier J-V, Jia Z, Nait Oumesmar B, et al. 2017. The intellectual disability protein PAK3 regulates oligodendrocyte precursor cell differentiation. *Neurobiology of Disease* **98**: 137–148.

Magno L, Lessard CB, Martins M, Lang V, Cruz P, Asi Y, Katan M, Bilsland J, Lashley T, Chakrabarty P, et al. 2019a. Alzheimer's disease phospholipase C-gamma-2 (PLCG2) protective variant is a functional hypermorph. *Alzheimer's Research & Therapy* **11**: 16. https://doi.org/10.1186/s13195-019-0469-0 (Accessed June 12, 2021).

Magno L, Lessard CB, Martins M, Lang V, Cruz P, Asi Y, Katan M, Bilsland J, Lashley T, Chakrabarty P, et al. 2019b. Alzheimer's disease phospholipase C-gamma-2 (PLCG2) protective variant is a functional hypermorph. *Alzheimer's Research & Therapy* **11**: 16.

Mao Y, Fisher DW, Yang S, Keszycki RM, Dong H. 2020. Protein-protein interactions underlying the behavioral and psychological symptoms of dementia (BPSD) and Alzheimer's disease. *PLOS ONE* **15**: e0226021.

Mapstone M, Gross TJ, Macciardi F, Cheema AK, Petersen M, Head E, Handen BL, Klunk WE, Christian BT, Silverman W, et al. 2020. Metabolic correlates of prevalent mild cognitive impairment and Alzheimer's disease in adults with Down syndrome. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **12**: e12028.

Marshe et. al, 2021. Genome-wide analysis suggests the importance of vascular processes and neuroinflammation in late-life antidepressant response | Translational Psychiatry. https://www.nature.com/articles/s41398-021-01248-3 (Accessed November 7, 2023).

Martin V, Zhuang F, Zhang Y, Pinheiro K, Gordân R. 2023. High-throughput data and modeling reveal insights into the mechanisms of cooperative DNA-binding by transcription factor proteins. *Nucleic Acids Research* **51**: 11600–11612.

Martini R, Schachner M. 1986. Immunoelectron microscopic localization of neural cell adhesion molecules (L1, N-CAM, and MAG) and their shared carbohydrate epitope and myelin basic protein in developing sciatic nerve. *J Cell Biol* **103**: 2439–2448.

Mathew B, Bathla S, Williams KR, Nairn AC. 2022. Deciphering Spatial Protein–Protein Interactions in Brain Using Proximity Labeling. *Molecular & Cellular Proteomics* **21**. https://www.mcponline.org/article/S1535-9476(22)00230-4/abstract (Accessed August 23, 2023).

Mathys et. al 2019. Single-cell transcriptomic analysis of Alzheimer's disease | Nature. https://www.nature.com/articles/s41586-019-1195-2 (Accessed March 27, 2023).

Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, Hayashi T, Nikaido I. 2017. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33**: 2314–2321.

Maurano et. al 2012. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA | Science. https://www.science.org/doi/10.1126/science.1222794 (Accessed June 11, 2024).

Mayran A, Drouin J. 2018. Pioneer transcription factors shape the epigenetic landscape. *Journal of Biological Chemistry* **293**: 13795–13804.

McCalla SG, Fotuhi Siahpirani A, Li J, Pyne S, Stone M, Periyasamy V, Shin J, Roy S. 2023. Identifying strengths and weaknesses of methods for computational network inference from single-cell RNA-seq data. *G3 Genes|Genomes|Genetics* **13**: jkad004.

Miao Z, Kim J. 2022. *Is single nucleus ATAC-seq accessibility a qualitative or quantitative measurement?* Bioinformatics http://biorxiv.org/lookup/doi/10.1101/2022.04.20.488960 (Accessed November 28, 2023).

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. 2013. Distributed Representations of Words and Phrases and their Compositionality. http://arxiv.org/abs/1310.4546 (Accessed June 30, 2024).

Mirny LA. 2010. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences* **107**: 22534–22539.

Mirsaeidi M, Gidfar S, Vu A, Schraufnagel D. 2016. Annexins family: insights into their functions and potential role in pathogenesis of sarcoidosis. *Journal of Translational Medicine* **14**: 89.

Mizuno et. al 2012. AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease | BMC Systems Biology | Full Text. https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-6-52 (Accessed November 7, 2023).

Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, Aerts S. 2019. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**: 2159–2161.

Morabito et. al 2020. Integrative genomics approach identifies conserved transcriptomic networks in Alzheimer's disease | Human Molecular Genetics | Oxford Academic. https://academic.oup.com/hmg/article/29/17/2899/5892988?login=true (Accessed November 7, 2023).

Morabito S, Miyoshi E, Michael N, Shahin S, Martini AC, Head E, Silva J, Leavy K, Perez-Rosendahl M, Swarup V. 2021. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat Genet* **53**: 1143–1155.

Morgunova E, Taipale J. 2017. Structural perspective of cooperative transcription factor binding. *Current Opinion in Structural Biology* **47**: 1–8.

Mukherjee S, Speed TP. 2008. Network inference using informative priors. *Proceedings of the National Academy of Sciences* **105**: 14313–14318.

Muraoka et. al, 2021. Enrichment of Neurodegenerative Microglia Signature in Brain-Derived Extracellular Vesicles Isolated from Alzheimer's Disease Mouse Models | Journal of Proteome Research. https://pubs.acs.org/doi/10.1021/acs.jproteome.0c00934 (Accessed November 7, 2023).

Murtaza N, Cheng AA, Brown CO, Meka DP, Hong S, Uy JA, El-Hajjar J, Pipko N, Unda BK, Schwanke B, et al. 2022. Neuron-specific protein network mapping of autism risk genes identifies shared biological mechanisms and disease-relevant pathologies. *Cell Reports* **41**: 111678.

Nagamine N, Kawada Y, Sakakibara Y. 2005. Identifying cooperative transcriptional regulations using protein–protein interactions. *Nucleic Acids Research* **33**: 4828–4837.

Nania et. al 2021. One-Third of COVID-19 Survivors Develop Brain Disorders. *AARP*. https://www.aarp.org/health/conditions-treatments/info-2021/brain-disorders-in-covid-survivors.html (Accessed November 10, 2023).

Nativio et. al 2020. An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease | Nature Genetics. https://www.nature.com/articles/s41588-020-0696-0 (Accessed November 7, 2023).

Naughton SX, Raval U, Pasinetti GM. 2020. Potential Novel Role of COVID-19 in Alzheimer's Disease and Preventative Mitigation Strategies. *Journal of Alzheimer's Disease* **76**: 21–25.

Nemani K, Li C, Olfson M, Blessing EM, Razavian N, Chen J, Petkova E, Goff DC. 2021. Association of Psychiatric Disorders With Mortality Among Patients With COVID-19. *JAMA Psychiatry* **78**: 380–386.

Nguyen H, Tran D, Tran B, Pehlivan B, Nguyen T. 2020. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. *Brief Bioinform* **22**: bbaa190.

Nica AC, Dermitzakis ET. 2013. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci* **368**: 20120362.

Nicodemus KK, Malley JD. 2009. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics* **25**: 1884–1890.

Nie J, Stewart R, Zhang H, Thomson JA, Ruan F, Cui X, Wei H. 2011. TF-Cluster: A pipeline for identifying functionally coordinated transcription factors via network decomposition of the shared coexpression connectivity matrix (SCCM). *BMC Systems Biology* **5**: 53.

Nie Y, Shu C, Sun X. 2020. Cooperative binding of transcription factors in the human genome. *Genomics* **112**: 3427–3434.

NIMH NIH et. al. NIMH » Mental Illness. https://www.nimh.nih.gov/health/statistics/mental-illness (Accessed November 10, 2023).

NINDS NIH et. al 2013. Brain Basics | National Institute of Neurological Disorders and Stroke. https://www.ninds.nih.gov/health-information/public-education/brain-basics (Accessed November 10, 2023).

Novikova G, Andrews SJ, Renton AE, Marcora E. 2021a. Beyond association: successes and challenges in linking non-coding genetic variation to functional consequences that modulate Alzheimer's disease risk. *Molecular Neurodegeneration* **16**: 27.

Novikova G, Kapoor M, Tcw J, Abud EM, Efthymiou AG, Chen SX, Cheng H, Fullard JF, Bendl J, Liu Y, et al. 2021b. Integration of Alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes. *Nat Commun* **12**: 1610.

NYU Langone Health, 2021. Schizophrenia Second Only to Age as Greatest Risk Factor for COVID-19 Death. *NYU Langone News*. https://nyulangone.org/news/schizophrenia-second-only-age-greatest-risk-factor-covid-19-death (Accessed November 7, 2023).

Ong I, O'CONNOR D, Shelef M, HEFFRON A, Baker D, AMJADI M, MCILWAIN S, KHULLAR S. 2024. Identification of sars-cov-2 epitopes discriminating covid-19 infection from control and methods of use. https://patents.google.com/patent/US20240044895A1/en (Accessed June 23, 2024).

Oughtred R, Rust J, Chang C, Breitkreutz B, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, et al. 2021. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* **30**: 187–200.

Overmyer et. al, 2021. Large-Scale Multi-omic Analysis of COVID-19 Severity - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2405471220303719?via%3Dihub (Accessed November 7, 2023).

Padi M, Quackenbush J. 2015. Integrating transcriptional and protein interaction networks to prioritize condition-specific master regulators. *BMC Syst Biol* **9**: 80.

Pairo-Castineira et. al 2020. Genetic mechanisms of critical illness in COVID-19 | Nature. https://www.nature.com/articles/s41586-020-03065-y (Accessed November 7, 2023).

Parab L, Pal S, Dhar R. 2022. Transcription factor binding process is the primary driver of noise in gene expression. *PLoS Genet* **18**: e1010535.

Patel et. al 2021. Cell-type-specific expression quantitative trait loci associated with Alzheimer disease in blood and brain tissue | Translational Psychiatry. https://www.nature.com/articles/s41398-021-01373-z (Accessed November 7, 2023).

Pearl JR, Colantuoni C, Bergey DE, Funk CC, Shannon P, Basu B, Casella AM, Oshone RT, Hood L, Price ND, et al. 2019. Genome-Scale Transcriptional Regulatory Network Models of Psychiatric and Neurodegenerative Disorders. *Cell Systems* **8**: 122-135.e7.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**: 2825–2830.

Peng X, Wang J, Peng W, Wu F-X, Pan Y. 2017. Protein–protein interactions: detection, reliability assessment and applications. *Briefings in Bioinformatics* **18**: 798–819.

Perna S, Pinoli P, Ceri S, Wong L. 2020. NAUTICA: classifying transcription factor interactions by positional and protein-protein interaction information. *Biology Direct* **15**: 13.

Perozzi B, Al-Rfou R, Skiena S. 2014. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710 http://arxiv.org/abs/1403.6652 (Accessed June 30, 2024).

Perry et. al, 2018. Religious Orders Study and Rush Memory and Aging Project - IOS Press. https://content.iospress.com/articles/journal-of-alzheimers-disease/jad179939 (Accessed November 7, 2023).

Poitelon et. al 2016. YAP and TAZ control peripheral myelination and the expression of laminin receptors in Schwann cells | Nature Neuroscience. https://www.nature.com/articles/nn.4316 (Accessed September 7, 2023).

Pratapa et. al 2020. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data | Nature Methods. https://www.nature.com/articles/s41592-019-0690-6 (Accessed March 27, 2023).

Project TB, 2015. hgu133acdf. *Bioconductor*. http://bioconductor.org/packages/hgu133acdf/ (Accessed November 7, 2023).

PsychENCODE Consortium. Resource.PsychEncode. http://resource.psychencode.org/ (Accessed November 7, 2023).

Qiao Y, Wang X-M, Mannan R, Pitchiaya S, Zhang Y, Wotring JW, Xiao L, Robinson DR, Wu Y-M, Tien JC-Y, et al. 2021. Targeting transcriptional regulation of SARS-CoV-2 entry factors ACE2 and TMPRSS2. *Proceedings of the National Academy of Sciences* **118**: e2021450118.

Qiu X, Rahimzamani A, Wang L, Ren B, Mao Q, Durham T, McFaline-Figueroa JL, Saunders L, Trapnell C, Kannan S. 2020. Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. *cels* **10**: 265-274.e11.

Quintela-López T, Ortiz-Sanz C, Serrano-Regal MP, Gaminde-Blasco A, Valero J, Baleriola J, Sánchez-Gómez MV, Matute C, Alberdi E. 2019. Aβ oligomers promote oligodendrocyte differentiation and maturation via integrin β1 and Fyn kinase signaling. *Cell Death Dis* **10**: 445.

R Core Team, 2021. R: The R Project for Statistical Computing. https://www.r-project.org/ (Accessed August 25, 2023).

Rabinovici GD. 2019. Late-onset Alzheimer Disease. *CONTINUUM: Lifelong Learning in Neurology* **25**: 14.

Rao S, Ahmad K, Ramachandran S. 2021. Cooperative binding between distant transcription factors is a hallmark of active enhancers. *Molecular Cell* **81**: 1651-1665.e4.

Reiken et. al 2022. Alzheimer's-like signaling in brains of COVID-19 patients - Reiken - 2022 - Alzheimer's & Dementia - Wiley Online Library. https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.12558 (Accessed November 7, 2023).

Reimand J, Kull M, Peterson H, Hansen J, Vilo J. 2007. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* **35**: W193–W200.

Reusch et. al, 2021. Frontiers | Neutrophils in COVID-19. https://www.frontiersin.org/articles/10.3389/fimmu.2021.652470/full (Accessed November 7, 2023).

Riddle R. 2012. Explaining “Alzheimer’s pathology” and Braak staging in “Alzheimer’s Disease.” *Brain Support Network*. https://www.brainsupportnetwork.org/explaining-alzheimers-pathology-and-braak-staging-in-alzheimers-disease/ (Accessed May 24, 2024).

Roberts NA. 2015. The Immune System. *Yale J Biol Med* **88**: 99.

Rödelsperger C, Köhler S, Schulz MH, Manke T, Bauer S, Robinson PN. 2009. Short ultraconserved promoter regions delineate a class of preferentially expressed alternatively spliced transcripts. *Genomics* **94**: 308–316.

Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, Ma’ayan A. 2016. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**: baw100.

Roussel J, Larcher R, Sicard P, Bideaux P, Richard S, Marmigère F, Thireau J. 2022. The autism-associated Meis2 gene is necessary for cardiac baroreflex regulation in mice. *Sci Rep* **12**: 20150.

Roy et. al 2020. JCI - Type I interferon response drives neuroinflammation and synapse loss in Alzheimer disease. https://www.jci.org/articles/view/133737 (Accessed November 7, 2023a).

Roy et. al 2020. PoLoBag: Polynomial Lasso Bagging for signed gene regulatory network inference from expression data | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/36/21/5187/5875056 (Accessed November 17, 2023b).

Roy S, Lagree S, Hou Z, Thomson JA, Stewart R, Gasch AP. 2013. Integrated Module and Gene-Specific Regulatory Inference Implicates Upstream Signaling Networks. *PLOS Computational Biology* **9**: e1003252.

Rustenhoven J, Smith AM, Smyth LC, Jansson D, Scotter EL, Swanson MEV, Aalderink M, Coppieters N, Narayan P, Handley R, et al. 2018a. PU.1 regulates Alzheimer’s disease-associated genes in primary human microglia. *Molecular Neurodegeneration* **13**: 44.

Rustenhoven J, Smith AM, Smyth LC, Jansson D, Scotter EL, Swanson MEV, Aalderink M, Coppieters N, Narayan P, Handley R, et al. 2018b. PU.1 regulates Alzheimer’s disease-associated genes in primary human microglia. *Molecular Neurodegeneration* **13**: 44.

Ryan and Petanceska et. al, 2022. AD Knowledge Portal. https://adknowledgeportal.synapse.org/ (Accessed August 22, 2023).

Safieh et. al, 2019. ApoE4: an emerging therapeutic target for Alzheimer’s disease | BMC Medicine | Full Text. https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1299-4 (Accessed November 7, 2023).

SaniyaKhullar 2024. SaniyaKhullar/Supplementary_Chapters_Dissertation: Supplementary Materials (Chapters A and B) for Saniya Khullar’s Dissertation. *GitHub Supplementary Materials (Chapters A and B) for Saniya Khullar’s Dissertation*.

https://github.com/SaniyaKhullar/Supplementary_Chapters_Dissertation/tree/main (Accessed July 22, 2024).

Satija et. al. 2024. Analysis, visualization, and integration of spatial datasets with Seurat. https://satijalab.org/seurat/articles/spatial_vignette (Accessed February 4, 2024).

Schep A, University S. 2023. motifmatchr: Fast Motif Matching in R. https://bioconductor.org/packages/motifmatchr/ (Accessed November 7, 2023).

Sealfon et. al 2021. Machine learning methods to model multicellular complexity and tissue specificity | Nature Reviews Materials. https://www.nature.com/articles/s41578-021-00339-3 (Accessed November 10, 2023).

Sealfon RSG, Wong AK, Troyanskaya OG. 2021. Machine learning methods to model multicellular complexity and tissue specificity. *Nat Rev Mater* **6**: 717–729.

Sevimoglu T, Arga KY. 2014. The role of protein interaction networks in systems biomedicine. *Comput Struct Biotechnol J* **11**: 22–27.

Shalek et. al 2014. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation | Nature. https://www.nature.com/articles/nature13437 (Accessed April 4, 2024).

Sharov AA, Nakatake Y, Wang W. 2022. Atlas of regulated target genes of transcription factors (ART-TF) in human ES cells. *BMC Bioinformatics* **23**: 377.

Shen M, Sirois CL, Guo Y, Li M, Dong Q, Méndez-Albelo NM, Gao Y, Khullar S, Kissel L, Sandoval SO, et al. 2023. Species-specific FMRP regulation of RACK1 is critical for prenatal cortical development. *Neuron* **111**: 3988-4005.e11.

Shojaie A, Michailidis G. 2009. Analysis of Gene Sets Based on the Underlying Regulatory Network. *J Comput Biol* **16**: 407–426.

Singh KV, Vig L. 2017. Improved prediction of missing protein interactome links via anomaly detection. *Appl Netw Sci* **2**: 2.

Sinha et. al 2020. Behavior-related gene regulatory networks: A new level of organization in the brain | PNAS. https://www.pnas.org/doi/10.1073/pnas.1921625117 (Accessed November 10, 2023).

Sinha KK, Bilokapic S, Du Y, Malik D, Halic M. 2023. Histone modifications regulate pioneer transcription factor cooperativity. *Nature* **619**: 378–384.

Skene PJ, Henikoff S. 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites ed. D. Reinberg. *eLife* **6**: e21856.

Skok Gibbs C, Jackson CA, Saldi G-A, Tjärnberg A, Shah A, Watters A, De Veaux N, Tchourine K, Yi R, Hamamsy T, et al. 2022. High-performance single-cell gene regulatory network inference at scale: the Inferelator 3.0. *Bioinformatics* **38**: 2519–2528.

Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT, et al. 2016. ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**: D726–D732.

Smith BN, Topp SD, Fallini C, Shibata H, Chen H-J, Troakes C, King A, Ticozzi N, Kenna KP, Soragia-Gkazi A, et al. 2017. Mutations in the vesicular trafficking protein annexin A11 are associated with amyotrophic lateral sclerosis. *Sci Transl Med* **9**: eaad9157.

Sönmezer C, Kleinendorst R, Imanci D, Barzaghi G, Villacorta L, Schübeler D, Benes V, Molina N, Krebs AR. 2021. Molecular co-occupancy identifies transcription factor binding cooperativity in vivo. *Mol Cell* **81**: 255-267.e6.

Sørensen et. al 2018. Biochemical Variables are Predictive for Patient Survival after Surgery for Skeletal Metastasis. A Prediction Model Development and External Validation Study. https://openorthopaedicsjournal.com/VOLUME/12/PAGE/469/ (Accessed July 19, 2024).

Specht AT, Li J. 2017. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics* **33**: 764–766.

Spitz F, Furlong EEM. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**: 613–626.

Srinivasan K, Friedman BA, Etxeberria A, Huntley MA, Brug MP van der, Foreman O, Paw JS, Modrusan Z, Beach TG, Serrano GE, et al. 2020. Alzheimer's Patient Microglia Exhibit Enhanced Aging and Unique Transcriptional Activation. *Cell Reports* **31**. https://www.cell.com/cell-reports/abstract/S2211-1247(20)30824-X (Accessed May 27, 2024).

Srinivasan R, Sun G, Keles S, Jones EA, Jang S-W, Krueger C, Moran JJ, Svaren J. 2012. Genome-wide analysis of EGR2/SOX10 binding in myelinating peripheral nerve. *Nucleic Acids Research* **40**: 6449–6460.

Srivastava D, Mahony S. 2020. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *Biochim Biophys Acta Gene Regul Mech* **1863**: 194443.

Stevanovic M, Drakulic D, Lazic A, Ninkovic DS, Schwirtlich M, Mojsin M. 2021. SOX Transcription Factors as Important Regulators of Neuronal and Glial Differentiation During Nervous System Development and Adult Neurogenesis. *Front Mol Neurosci* **14**: 654031.

Stratton JA, Kumar R, Sinha S, Shah P, Stykel M, Shapira Y, Midha R, Biernaskie J. 2017. Purification and Characterization of Schwann Cells from Adult Human Skin and Nerve. *eNeuro* **4**. https://www.eneuro.org/content/4/3/ENEURO.0307-16.2017 (Accessed October 18, 2023).

Strauss MJ, Niederkrotenthaler T, Thurner S, Kautzky-Willer A, Klimek P. 2021. Data-driven identification of complex disease phenotypes. *Journal of The Royal Society Interface* **18**: 20201040.

Stuart et. al 2021. Single-cell chromatin state analysis with Signac | Nature Methods. https://www.nature.com/articles/s41592-021-01282-5 (Accessed March 27, 2023).

Sturm VE, Haase CM, Levenson RW. 2016. Chapter 22 - Emotional Dysfunction in Psychopathology and Neuropathology: Neural and Genetic Pathways. In *Genomics, Circuits, and Pathways in Clinical Neuropsychiatry* (eds. T. Lehner, B.L. Miller, and M.W. State), pp. 345–364, Academic Press, San Diego https://www.sciencedirect.com/science/article/pii/B9780128001059000226 (Accessed May 24, 2024).

Su et. al 2021. Activation of NF-κB and induction of proinflammatory cytokine expressions mediated by ORF7a protein of SARS-CoV-2 | Scientific Reports. https://www.nature.com/articles/s41598-021-92941-2 (Accessed November 7, 2023).

Su L, Liu G, Guo Y, Zhang X, Zhu X, Wang J. 2022. Integration of Protein-Protein Interaction Networks and Gene Expression Profiles Helps Detect Pancreatic Adenocarcinoma Candidate Genes. *Frontiers in Genetics* **13**. https://www.frontiersin.org/articles/10.3389/fgene.2022.854661 (Accessed November 15, 2023).

Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, et al. 2023. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research* **51**: D638–D646.

Tajouri L, Fernandez F, Griffiths LR. 2007. Gene Expression Studies in Multiple Sclerosis. *Curr Genomics* **8**: 181–189.

Tan G, Lenhard B. 2016. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32**: 1555–1556.

Tao et. al 2019. Variation in SIPA1L2 is correlated with phenotype modification in Charcot– Marie–Tooth disease type 1A - Tao - 2019 - Annals of Neurology - Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1002/ana.25426 (Accessed July 14, 2023).

Team BC, Maintainer BP, et. al 2019. TxDb.Hsapiens.UCSC.hg38.knownGene. *Bioconductor*. http://bioconductor.org/packages/TxDb.Hsapiens.UCSC.hg38.knownGene/ (Accessed August 25, 2023).

THE GTEX CONSORTIUM. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**: 1318–1330.

Thompson et. al 2015. Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution | Annual Review of Cell and Developmental Biology. https://www.annualreviews.org/doi/10.1146/annurev-cellbio-100913-012908?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0pubmed (Accessed November 10, 2023).

Thompson et. al 2022. COVID Appears to Raise Risk for Alzheimer's Disease. *US News & World Report*. //www.usnews.com/news/health-news/articles/2022-09-16/covid-appears-to-raise-risk-for-alzheimers-disease (Accessed November 7, 2023).

Titsias MK, Honkela A, Lawrence ND, Rattray M. 2012. Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison. *BMC Systems Biology* **6**: 53.

Tran et. al 2019. Defining Reprogramming Checkpoints from Single-Cell Analyses of Induced Pluripotency - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2211124719305297 (Accessed March 9, 2024).

Tremblay et. al, 2020. Frontiers | Neuropathobiology of COVID-19: The Role for Glia. https://www.frontiersin.org/articles/10.3389/fncel.2020.592214/full (Accessed November 7, 2023).

Turley et al. Executive Summary | Pan UKBB. https://pan-dev.ukbb.broadinstitute.org/docs/summary (Accessed April 10, 2024).

Van de Sande et. al, 2020. A scalable SCENIC workflow for single-cell gene regulatory network analysis | Nature Protocols. https://www.nature.com/articles/s41596-020-0336-2 (Accessed November 10, 2023).

van der Wijst M, de Vries D, Groot H, Trynka G, Hon C, Bonder M, Stegle O, Nawijn M, Idaghdour Y, van der Harst P, et al. 2020. The single-cell eQTLGen consortium eds. H. Pérez Valle, P. Rodgers, S.B. Montgomery, and M. Fagny. *eLife* **9**: e52155.

van Dijk et. al 2018. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion: Cell. https://www.cell.com/cell/fulltext/S0092-8674(18)30724-4?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867418307244%3Fshowall%3Dtrue (Accessed August 23, 2024).

van Duin L, Krautz R, Rennie S, Andersson R. 2023. Transcription factor expression is the main determinant of variability in gene co-activity. *Molecular Systems Biology* **19**: e11392.

Vargas DM, De Bastiani MA, Zimmer ER, Klamt F. 2018. Alzheimer's disease master regulators analysis: search for potential molecular targets and drug repositioning candidates. *Alzheimer's Research & Therapy* **10**: 59.

Vasile et. al, 2017. Human astrocytes: structure and functions in the healthy brain | Brain Structure and Function. https://link.springer.com/article/10.1007/s00429-017-1383-5 (Accessed November 7, 2023).

Venkatesan K, Rual J-F, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh K-I, et al. 2009. An empirical framework for binary interactome mapping. *Nat Methods* **6**: 83–90.

Vickers AJ, Elkin EB. 2006. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* **26**: 565–574.

Wallace C. 2020. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLOS Genetics* **16**: e1008720.

Wan Q, Tang J, Han Y, Wang D. 2018. Co-expression modules construction by WGCNA and identify potential prognostic markers of uveal melanoma. *Experimental Eye Research* **166**: 13–20.

Wang AR, Khullar S, Brown J, Baschnagel A, Buehler D, Kendziorski C, Iyer G. 2022a. Abstract 3859: Remodeling the extracellular matrix environment enables the dissemination of primary tumor cells through a chemokine gradient to establish brain metastasis in non-small cell lung cancer adenocarcinoma. *Cancer Research* **82**: 3859.

Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, Clarke D, Gu M, Emani P, Yang YT, et al. 2018a. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**: eaat8464.

Wang et. al 2018. Comprehensive functional genomic resource and integrative model for the human brain | Science. https://www.science.org/doi/10.1126/science.aat8464?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed (Accessed March 27, 2023).

Wang H, Huang B, Wang J. 2021a. Predict long-range enhancer regulation based on protein–protein interactions between transcription factors. *Nucleic Acids Res* **49**: 10347–10368.

Wang J, Liu Q, Sun J, Shyr Y. 2016. Disrupted cooperation between transcription factors across diverse cancer types. *BMC Genomics* **17**: 560.

Wang M, Zhang L, Gage FH. 2019. Microglia, complement and schizophrenia. *Nat Neurosci* **22**: 333–334.

Wang Q, Guo M, Chen J, Duan R. 2023. A gene regulatory network inference model based on pseudo-siamese network. *BMC Bioinformatics* **24**: 163.

Wang T, Bai J, Nabavi S. 2021b. Single-cell classification using graph convolutional networks. *BMC Bioinformatics* **22**: 364.

Wang Y, Duan X, Zhou X, Wang R, Zhang X, Cao Z, Wang X, Zhou Z, Sun Y, Peng D. 2022b. ANXA11 mutations are associated with amyotrophic lateral sclerosis–frontotemporal dementia. *Front Neurol* **13**: 886887.

Wang Y, Hu X, Jiang X, He T, Yuan J. 2015. Predicting microbial interactions by using network-constrained regularization incorporating covariate coefficients and connection signs. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 635–638.

Wang Y, Shan Q, Pan J, Yi S. 2018b. Actin Cytoskeleton Affects Schwann Cell Migration and Peripheral Nerve Regeneration. *Front Physiol* **9**: 23.

Wang Y, Zhang X-S, Xia Y. 2009. Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data. *Nucleic Acids Research* **37**: 5943–5958.

Weber M, Striaukas J, Schumacher M, Binder H. Regularized regression when covariates are linked on a network: the 3CoSE algorithm. *J Appl Stat* **50**: 535–554.

Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**: 1113–1120.

Wightman et. al, 2021. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease | Nature Genetics. https://www.nature.com/articles/s41588-021-00921-z (Accessed November 7, 2023).

Winter D, Chamberlain S, Guangchun H. 2020. rentrez: "Entrez" in R. https://cran.r-project.org/web/packages/rentrez/index.html (Accessed November 7, 2023).

Wong AK, Krishnan A, Troyanskaya OG. 2018. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic Acids Research* **46**: W65–W70.

Wong et. al 2021. Decoding disease: from genomes to networks to phenotypes | Nature Reviews Genetics. https://www.nature.com/articles/s41576-021-00389-x (Accessed April 18, 2024).

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, et al. 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2**. https://www.cell.com/the-innovation/abstract/S2666-6758(21)00066-7 (Accessed August 25, 2023).

Xiong N, Schiller MR, Li J, Chen X, Lin Z. 2021. Severe COVID-19 in Alzheimer's disease: APOE4's fault again? *Alzheimers Res Ther* **13**: 111.

Yang P, Li X, Chua H-N, Kwoh C-K, Ng S-K. 2014. Ensemble Positive Unlabeled Learning for Disease Gene Identification. *PLOS ONE* **9**: e97079.

Yashiro T, Nakano S, Nomura K, Uchida Y, Kasakura K, Nishiyama C. 2019. A transcription factor PU.1 is critical for Ccl22 gene expression in dendritic cells and macrophages. *Sci Rep* **9**: 1161. http://www.nature.com/articles/s41598-018-37894-9 (Accessed April 19, 2021).

Yeger-Lotem E, Sharan R. 2015. Human protein interaction networks across tissues and diseases. *Front Genet* **6**: 257.

Yi M, Wei T, Wang Y, Lu Q, Chen G, Gao X, Geller HM, Chen H, Yu Z. 2017. The potassium channel KCa3.1 constitutes a pharmacological target for astrogliosis associated with ischemia stroke. *J Neuroinflammation* **14**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5644250/ (Accessed April 19, 2021).

Yim et. al 2022. Disentangling glial diversity in peripheral nerves at single-nuclei resolution | Nature Neuroscience. https://www.nature.com/articles/s41593-021-01005-1 (Accessed October 24, 2023).

Yu B, Chen C, Zhou H, Liu B, Ma Q. 2020. GTB-PPI: Predict Protein–protein Interactions Based on L1-regularized Logistic Regression and Gradient Tree Boosting. *Genomics, Proteomics & Bioinformatics* **18**: 582–592.

Yu D, Chojnowski G, Rosenthal M, Kosinski J. 2023. AlphaPulldown—a python package for protein–protein interaction screens using AlphaFold-Multimer. *Bioinformatics* **39**: btac749.

Yu G, Wang L-G, Han Y, He Q-Y. 2012. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS* **16**: 284–287.

Yu G, Wang L-G, He Q-Y. 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**: 2382–2383.

Zaborowski AB, Walther D. 2020. Determinants of correlated expression of transcription factors and their target genes. *Nucleic Acids Research* **48**: 11347–11369.

Zass LJ, Hart SA, Seedat S, Hemmings SMJ, Malan-Müller S. 2017. Neuroinflammatory genes associated with post-traumatic stress disorder: implications for comorbidity. *Psychiatric Genetics* **27**: 1.

Zeng B, Bendl J, Kosoy R, Fullard JF, Hoffman GE, Roussos P. 2022. Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits. *Nat Genet* **54**: 161–169.

Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, Qiu Y, Li YE, Gaulton KJ, Wang A, et al. 2021. A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**: 5985-6001.e19.

Zhang R, Ren Z, Chen W. 2018. SILGGM: An extensive R package for efficient statistical inference in large-scale gene networks. *PLOS Computational Biology* **14**: e1006369.

Zhang S, Pyne S, Pietrzak S, Halberg S, McCalla SG, Siahpirani AF, Sridharan R, Roy S. 2023. Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. *Nat Commun* **14**: 3064.

Zhang W, Wan Y, Allen GI, Pang K, Anderson ML, Liu Z. 2013. Molecular pathway identification using biological network-regularized logistic models. *BMC Genomics* **14**: S7.

Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, Luo T, Xu L, Liao G, Yan M, et al. 2019. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Research* **47**: D721–D728.

Zhang Z, Liu X, Zhang S, Song Z, Lu K, Yang W. 2024. A review and analysis of key biomarkers in Alzheimer's disease. *Front Neurosci* **18**. https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2024.1358998/full (Accessed June 4, 2024).

Zhao Y. 2023. TFSyntax: a database of transcription factors binding syntax in mammalian genomes. *Nucleic Acids Research* **51**: D306–D314.

Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. 2019. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**: 1523.

Zinati Y, Takiddeen A, Emad A. 2024. GRouNdGAN: GRN-guided simulation of single-cell RNA-seq data using causal generative adversarial networks. *Nat Commun* **15**: 4055.

Zou H, Hastie T. 2005. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**: 301–320.

An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders | Nature Genetics. https://www.nature.com/articles/s41588-018-0130-z (Accessed November 10, 2023a).

Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution - ScienceDirect.

https://www.sciencedirect.com/science/article/pii/S0092867411013511?via%3Dihub (Accessed November 14, 2023b).

Detection of cooperatively bound transcription factor pairs using ChIP-seq peak intensities and expectation maximization - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6049898/ (Accessed March 27, 2023c).

Estimating the size of the human interactome | PNAS. https://www.pnas.org/doi/10.1073/pnas.0708078105 (Accessed November 15, 2023d).

Know Thy Friends, know ThyselF – Transcription Factor (TF) regulation resolved | Springer Nature Protocols and Methods Community. https://protocolsmethods.springernature.com/posts/know-thy-friends-know-thyself-transcription-factor-tf-regulation-resolved (Accessed March 27, 2023e).

Noncoding RNA and Gene Expression | Learn Science at Scitable. https://www.nature.com/scitable/topicpage/regulation-of-transcription-and-gene-expression-in-1086/ (Accessed November 17, 2023f).

Sam Morabito | scWGCNA. https://smorabit.github.io/tutorials/9_scWGCNA_tutorial/ (Accessed November 14, 2023g).

Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology: Cell. https://www.cell.com/cell/fulltext/S0092-8674(23)00973-X?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867423009730X%3Fshowall%3Dtrue (Accessed November 14, 2023h).

Systematic Functional Annotation of Somatic Mutations in Cancer: Cancer Cell. https://www.cell.com/cancer-cell/fulltext/S1535-6108(18)30021-7 (Accessed November 10, 2023i).

Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders: Cell. https://www.cell.com/cell/fulltext/S0092-8674(15)00430-4?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867415004304%3Fshowall%3Dtrue (Accessed November 10, 2023j).