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

By: Saniya Khullar

Supplementary materials for the dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

(Biomedical Data Science)

UNIVERSITY OF WISCONSIN-MADISON

2024

Table of Contents

§ Chapter A: Supplemental Materials for SNPheno

§ A.1 Supplementary methods and materials

A.1.1 Building a Gene-Expression Based Gene Reguatory Network Step 3

Here, we explain further how we performed this GRN Step 3 to build our gene expression-based gene regulatory network (GRN). In GRN Step 3, we combined TF list (Lambert et al. 2018) with JASPAR's list (Fornes et al. 2020) to generate a final list of TFs, which we used to find candidate TFs for each brain region (based on respective gene expression data genes available). We use 3 different tools (and their corresponding algorithms on bulk RNA-seq data) to infer TF-TG gene regulatory pairs with strong expression relations. TreNA Ensemble Solver (systemsbiology.org> et al. 2023) with defaults (geneCutoff: 0.1; solvers: RandomForest, Spearman, Lasso, LassoPV, Ridge, Pearson) constructed a GRN linking TFs to TGs. GENIE3 (Huynh-Thu et. al, 2010) also predicted a GRN via Random Forest regression, predicting each gene's expression pattern from expression patterns of all TFs (retaining TF-TG pairs with weights > 0.0025). RTN predicts TFs to TGs by calculating Mutual Information between TFs and all genes; permutation analysis with 1,000 permutations, bootstrapping, and ARACNe algorithm selected most meaningful network edges for removing possible false positives and false negative network edges from prediction. TF-TG pairs found in at least 2 of these 3 tools comprised our gene expression-based network.

A.1.2 TF to statistically significantly regulated gene co-expression modules in the Hippocampus and Lateral Temporal Lobe (LTL) brain regions

In addition to predicting TFs for individual genes, we inferred TFs significantly co-regulating Hippo. and LTL modules via Master Regulatory Analysis (MRA) on the RTN-inferred network by the RTN package (Groeneveld et al. 2023). For each gene module, MRA performed enrichment analysis using the inferred GRN, phenotype (Module Membership correlation of all genes to that module), and hits (module genes). RTN then applied the hypergeometric test for overlaps between TFs and genes (using gene expression data) and found significant TFs for each module.

A.1.3 Decision Curve Analysis (DCA)

We used DCA (Vickers et al. 2019) to evaluate and compare the machine learning models of those brain-region AD-Covid genes and benchmark genes for predicting Covid severity. DCA is widely used for making medical decisions, improving upon traditional comparison metrics (e.g., AUROC) for predictive models and other approaches requiring additional information to address individuals' clinical consequences for individuals. Decision Curves show how the

Net Benefit of each model varies across probability thresholds. Given a model and a threshold probability pT, patients will be sent to the Intensive Care Unit (ICU) if their percentage risks for Covid severity (i.e., ICU) from the model are greater than or equal to pT. Based on this, the true positive (TP) count is the number of Covid-19 severe individuals sent to the ICU, and the False Positive (FP) is the number of Covid-19 non-ICU individuals sent to the ICU. Thus, pT inherently represents subjective clinician preferences for FPs versus False Negatives (FNs: number of severe Covid-19 patients who were wrongly not sent to the ICU). Based on TP, FP and pT, the DCA then calculates Net Benefit = $\frac{TP}{N} - \left(\left(\frac{FP}{N}\right) \times \frac{pT}{1-pT}\right)$, where N is the total number of patients (N=100 here). Thus, the Net Benefit represents the benefit of true positive ratio (TP/N) from false positive ratio (FP/N) weighted by odds of pT (i.e., pT/(1- pT)). DCA provides a simple, personalized risk-tolerance based approach of using pT to weight the FN and FP mistakes: lower thresholds represent a fear of FNs over FPs, and vice-versa. For instance, for a clinician who sends a Covid-19 positive individual with predicted severity of at least pT = 20% to the ICU, the utility of treating a Covid-19 severe individual is 4 times greater than the harm of needlessly sending a non-severe Covid-19 patient to the ICU. We compared our predictive models with 2 extremes: Treat All (predict 1 for all Covid-19 positive patients and send all to ICU regardless of severity) and Treat None (predict 0 for all positive patients and send none to ICU). Practically, a clinician ought to opt for the predictive model or extreme intervention strategy with the highest Net Benefit based on that clinician's preferred pT; thus, two clinicians (who may have their own, different pT values) may obtain different optimal results. Thus, DCA can evaluate the clinical usability of a Covid-19 severity prediction model based on its Net Benefit across clinically reasonable pT values. The threshold of each model with the highest Net Benefit corresponds to the optimal decision probability for sending Covid-19 patients to ICU or not, i.e., the "optimal" threshold. Note that since 50% of our Covid-19 positive patients are in the ICU, the maximum Net Benefit is 0.50. Finally, we performed DCA using code from Memorial Sloan Kettering Cancer Center. Besides Covid severity, we also calculated gene expression correlations with Covid-19 and non-Covid for the genes from the AD KEGG pathway for 3 brain regions.

A.1.4 Building and Analyzing SNP Regulatory Networks

This is our framework for SNP-effected GRNs:

For Alzheimer's Disease (AD) Single Nucleotide Polymorphisms (SNPs), we pooled together summary statistics from several Genome-Wide Association Studies (**Table A.2** and **File A8**). Thus, some SNPs could fall in multiple AD GWAS and in that case, we would use the most significant summary statistics (e.g. effect sizes) based on the

GWAS P value; that is, we used the summary statistic that was the most significant for that SNP (has the smallest p-value). In general, SNPs with a negative GWAS effect size for AD and/or severe Covid-19 are predicted as being 'protective' or helpful as they are associated with a decreased risk of AD and/or severe Covid-19, respectively. Similarly, SNPs with a positive effect size are considered 'harmful' given that they are associated with an increased risk of disease. When needed, we mapped SNPs in the hg19 chr#:basePair (chromosome #: base pair) position to reference SNP IDs (rsIDs) as MotifbreakR(Coetzee et al. 2015) works with mapped rsIDs. When we ran MotifbreakR package, we used the following parameters (based on hg19 or GRCh37 genome build): SNP database = SNPlocs.Hsapiens.dbSNP144.GRCh37 (Pagès 2017); BSgenome database = Bsgenome.Hsapiens.UCSC.hg19; methodology used (all 3 available): default, information content (ic), log; motif databases: HOCOMOCO, FactorBook, HOMER; threshold 0.001. MotifbreakR predicted TFs impacted by the given SNPs and returns predictions of the impact of the SNP on the TF Binding Site (TFBS). We used alleleDiff value to determine the SNP impact on a TF; SNP-TF interactions with a high absolute value for alleleDiff (i.e. |alleleDiff|) would be predicted as 'strong' effects of the SNP on TF binding (otherwise, the SNP effect was predicted as weak). An alleleDiff value > 0, predicts the SNP increases TF affinity for the TFBS based on the TF's sequence-specific motif; alleleDiff < 0 predicts the SNP disrupts TF binding. Ultimately, we link SNPs to our full GRNs across our brain regions. Thus, we linked SNPs to TGs from regulatory elements (REs) with altered TFBSs. To capture more relations, we added a buffer (extension): 10,000 base pairs to the start and end positions of enhancers and 2,000 base pairs to the start and end positions of promoters. We noted the closest distance from the SNP to the regulatory region start or end position if the SNP was not found within the RE. We ensured that the SNP that was linked to the TF-RE-TG (SNP found within the RE or the respective buffer region) was also associated with the TF (via motifbreakR: SNP-TF link).

Our gene co-expression network analysis (for each brain region) assigned each TG to 1 gene co-expression modules and determined biological enrichments for those gene modules. It further associated TGs and their gene modules with AD phenotypes. Thus, we ultimately could link our SNPs to our full GRNs across our brain regions. Our full GRNs had linked TFs to TFBSs on REs (like enhancers and promoters) to TGs and modules and phenotypes/enrichments. Our SNP-effected-GRN therefore linked AD and/or severe Covid-19 SNPs to TFs to REs to TGs and gene modules (and biological enrichments for modules). Those TGs and gene modules are associated with various AD phenotypes. Our SNP Regulatory Networks (SNP-effected-GRN) may be analyzed in many

different ways. Fundamentally, this network links non-coding SNPs (that are associated with AD and/or Covid-19 severity) to our predicted GRNs across the 3 brain regions. These are the nodes in our SNP-effected-GRN:

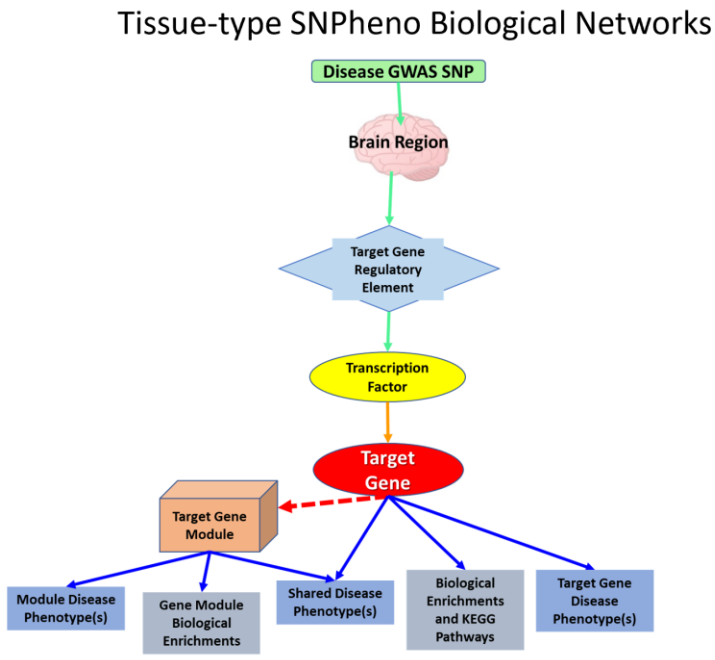
- SNPs associated with AD and/or severe Covid-19: We have reference SNP id and hg19 position.
- SNP p-value: significance of the association of that SNP with the condition (AD or Covid-19 severity) based on the GWAS. If the condition is AD and the SNP is found in several GWAS studies, we use the smallest p-value found for that SNP across those studies
- SNP effect size: it is based on the GWAS summary statistics (most significant for AD for a given SNP). When effect size is negative, the SNP is protective (associated with decreased risk of disease); when effect size is positive, the SNP is harmful (associated with higher risk of disease)
- REs like enhancers or promoters: chromosome #: regulatory region start – regulatory region end.
- TFs (Transcription Factors)
- Regulated or target genes (TGs)
- AlleleDiff: < 0 predicts the SNP decreases TF Binding; > 0 predicts SNP increases TF binding
- Distance of SNP from the Regulatory Element (RE): if this is 0 then the SNP falls within the RE. Otherwise, the SNP can be up to 10,000 base pairs away (i.e. 10 kilobase pairs (kbp)) from an enhancer start or end and up to 2,000 base pairs (i.e. 2 kbp) away from a promoter start or end.
- expression quantitative trait loci (eQTL) support: number of eQTL sources (tissues/cell-types for the brain/blood) that have found significant associations between the SNP and the TG expression

We can use information in **Files A.1** to **A.3** (provide information on gene modules for the TGs, gene enrichments for the modules, and associations between TGs/modules and AD phenotypes) to finish building out the full SNP-effected-GRN. Please note that we did not provide the TG modules, AD phenotypes, and module enrichments to Supplementary File 8 due to file size constraints. However, our webtool visualizes the entire SNP-effected-GRN. Please note that when looking at our SNP-effected-GRN, it may be helpful to prioritize the following:

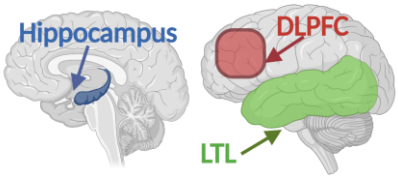
- TGs that belong to the list of flagged predictive, optimal AD-Covid genes. It may also be of interest to look at TFs that are optimal AD-Covid genes or at common TGs dysregulated by AD and severe Covid-19.
- SNP-effected-GRNs that have SNPs directly within the RE (distance of 0).
- SNP-effected-GRNs that have eQTL support with experimentally validated SNP-TG links
- Common AD and severe Covid-19 SNPs. (We found up to 14 in our GWAS datasets)

- SNPs with a significant p-value
- TGs that are positively correlated with AD phenotypes and/or belong to a 'phenotype-enriched' gene co-expression module (significantly positively associated with AD-related phenotypes)

Ultimately, there are many stories that can be uncovered by our SNP-effected-GRN. We hope that this network may help researchers understand better the role of neuroimmunology and other factors in AD, cognitive impairment associated with AD, mechanisms associated with Covid-19 severity, and AD-Covid links. For example, our SNP-effected-GRN predicts previously unknown SNPs and TFs associated with NFKB TF activation in AD, which may make NFKB TFs neuroprotective or neurotoxic. Tissue-type SNPheno Biological Networks represent our SNP-effected-GRNs in this case.



A.1.5 Regions and Applications



Region	Disease	# of Human Samples	Application
Hippocampus	Alzheimer's Disease (AD)	31	Build region-specific Gene Regulatory Networks (SNPs to Phenotypes) for AD
Lateral Temporal Lobe (LTL)		30	
Dorsolateral Prefrontal Cortex (DLPFC)		638	
Blood Serum	Covid-19 Severity	100	Identify potential AD neuroinflammatory risk genes

§ A.2 Supplementary figures

Figure A.1 Heatmaps for all gene co-expression modules in the Lateral Temporal Lobe (LTL) and Dorsolateral Prefrontal Cortex (DLPFC) brain regions. This is based on **Figure 2.2.**

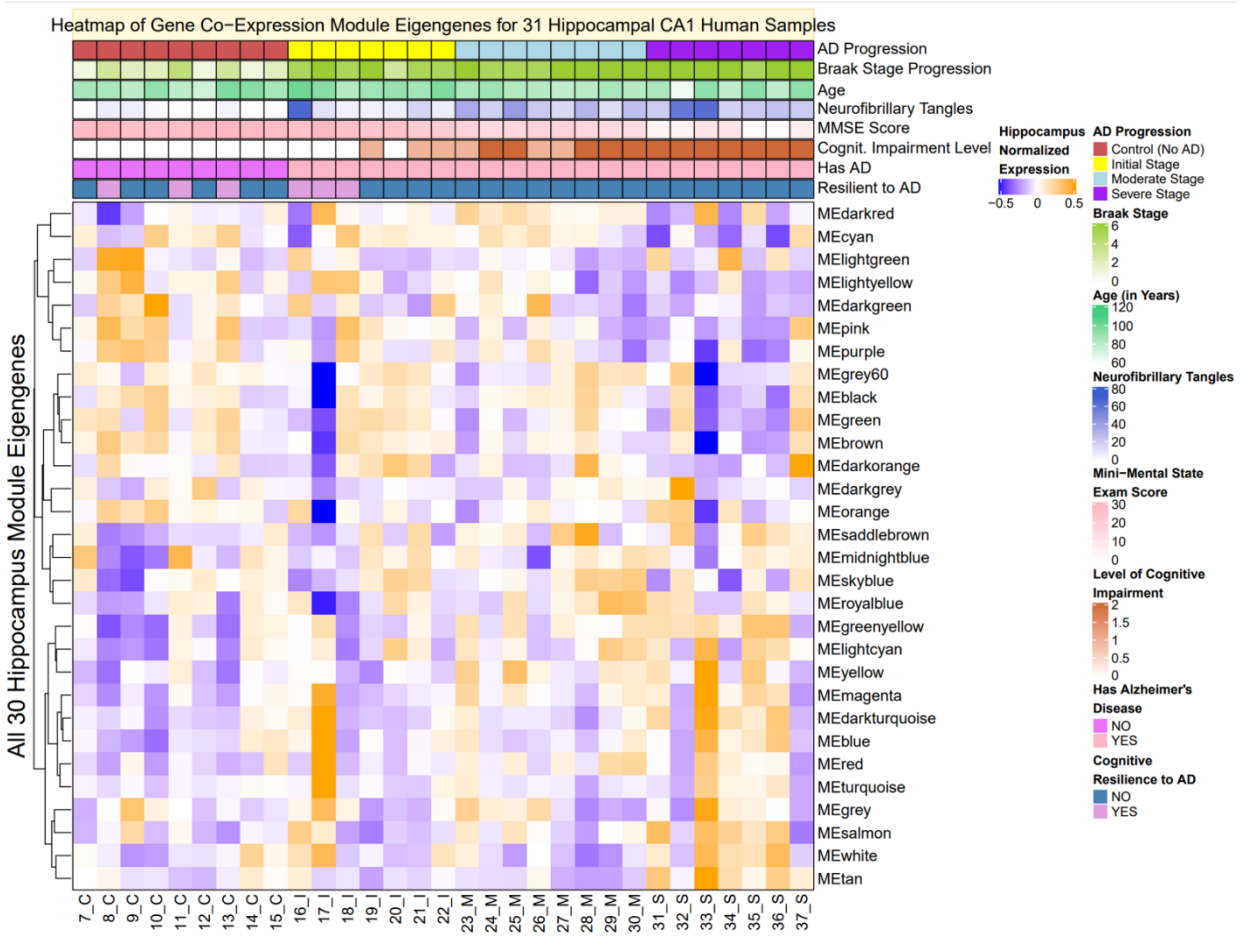


Figure A.1A) - Module eigengenes (MEs) of all 30 gene co-expression modules in the Hippocampus CA1 human samples.

Red: high expression level. Blue: low expression level. This heatmap presents the different expression dynamic patterns during AD progression for all 30 gene modules in the Hippocampus Ca1 region after Weighted Gene Co-Expression Network Analysis with k-means. This heatmap illustrates how the represented modular expression patterns (module eigengenes) vary for the AD phenotypes, such as AD Progression, Braak Stage Progression, Age, Neurofibrillary Tangles (NFTs), Mini-Mental State Exam Score, Level of Cognitive Impairment, Alzheimer's Disease, and Cognitive Resilience to AD.

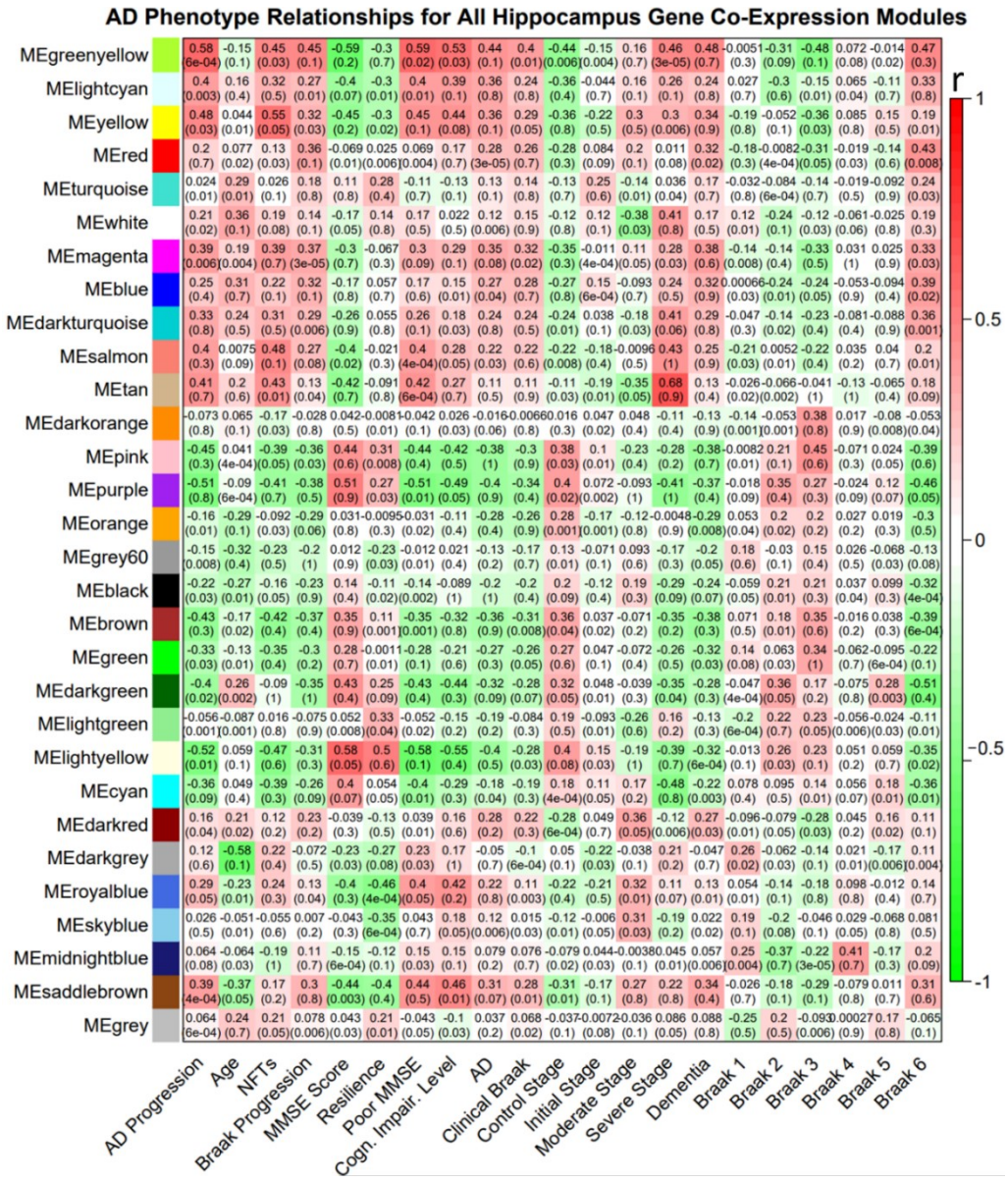


Figure A.1B) – AD phenotype relationships for all 30 Hippocampus Gene Co-expression modules.

This figure further links all 30 gene modules in the Hippocampus to select respective AD phenotypes; these heatmaps present the Pearson correlation (r) on top and the resulting p-value below, for each module-phenotype relationship. These phenotypes are AD Progression, Age, NFTs, Braak Progression, MMSE Score, Resilience, Poor MMSE, Cognitive Impairment Level, AD, Clinical Braak Stage, Control Stage, Initial Stage, Moderate Stage, Severe Stage, Dementia, Braak 1, Braak 2, Braak 3, Braak 4, Braak 5, Braak 6. For example, the Tan Module has a very strong and significant positive correlation (r = 0.68) with the Severe Stage.

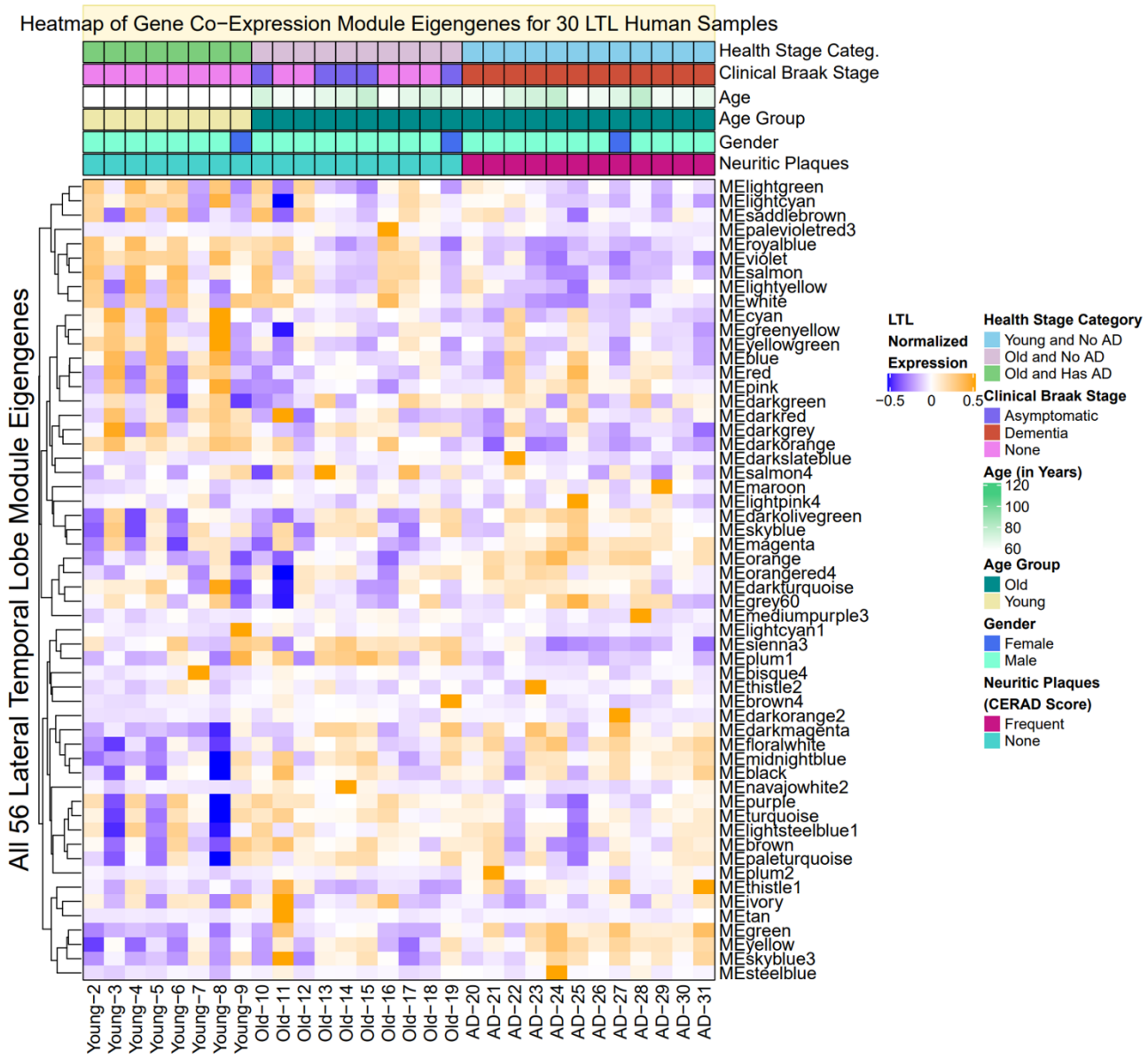


Figure A.1C) – Heatmap of All 56 Gene Co-expression Module Eigengenes for the 30 LTL Human Samples.

This heatmap presents the different expression dynamic patterns during AD progression for all 56 gene modules in the LTL region after Weighted Gene Co-Expression Network Analysis with k-means. This heatmap illustrates how the represented modular expression patterns (module eigengenes) vary for AD phenotypes, such as Health Category (Aging and Developing AD), Clinical Braak Stage, Age (in Years), Age Group, Gender, Neuritic Plaques (CERAD Score).

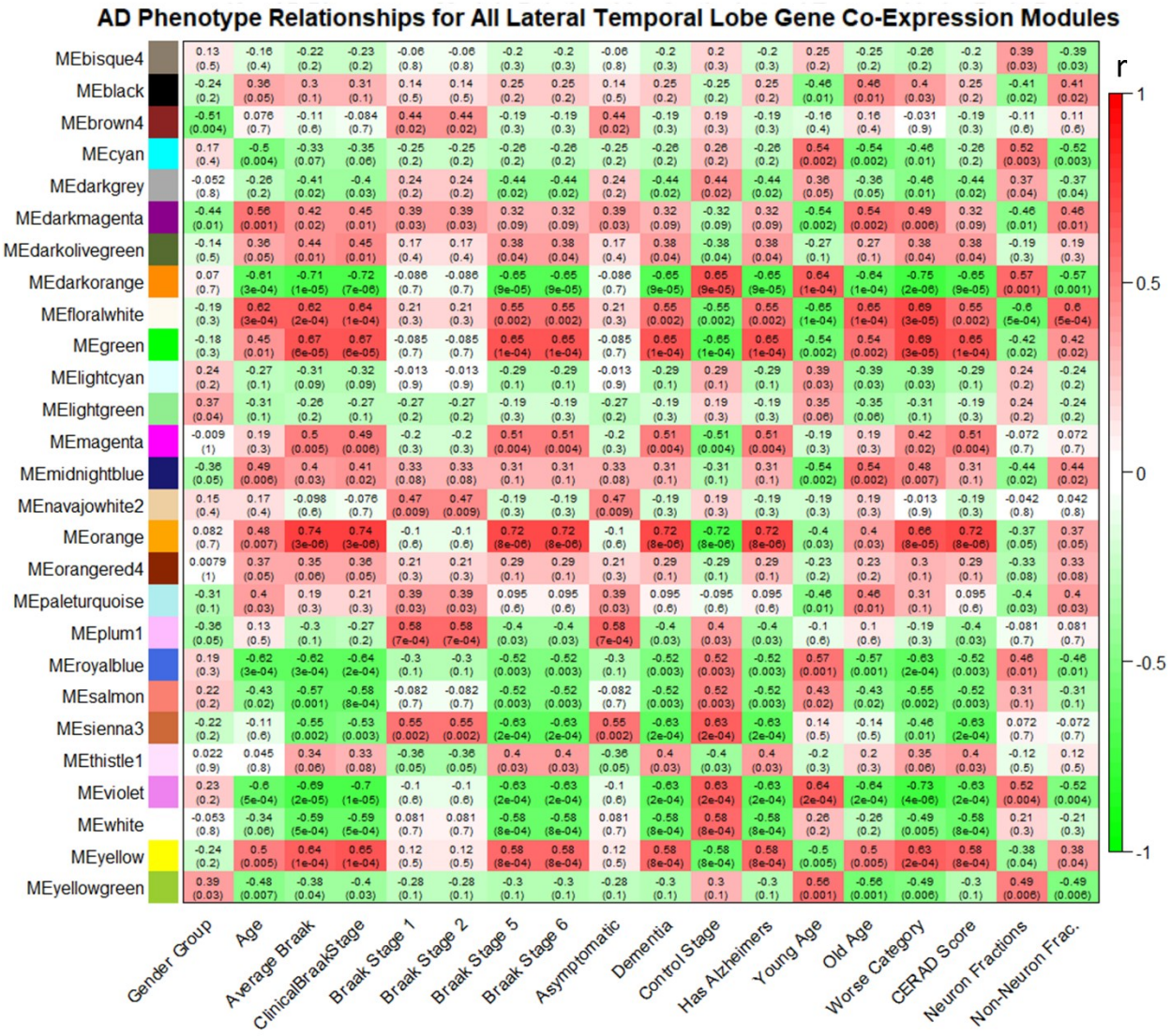


Figure A.1D) – AD Phenotype Relationships for All 56 Lateral Temporal Lobe Gene Co-Expression Modules.

This figure further links all 56 gene modules in the LTL to select respective AD phenotypes; these heatmaps present the Pearson correlation (r) on top and the resulting p-value below, for each module-phenotype relationship. These phenotypes are Gender Group, Age, Average Braak, Clinical Braak Stage, Braak 1, Braak 2, Braak 5, Braak 6, Asymptomatic (Braak 1 or 2), Dementia (Braak 5 or 6), Control Stage, AD (Alzheimer's Disease), Young Age, Old Age, Worse Category (Aging and Developing AD), CERAD Score, Neuron Fractions, Non-Neuron Fractions. For example, the Orange Module has a very strong and significant positive correlation (r = 0.72) with AD.

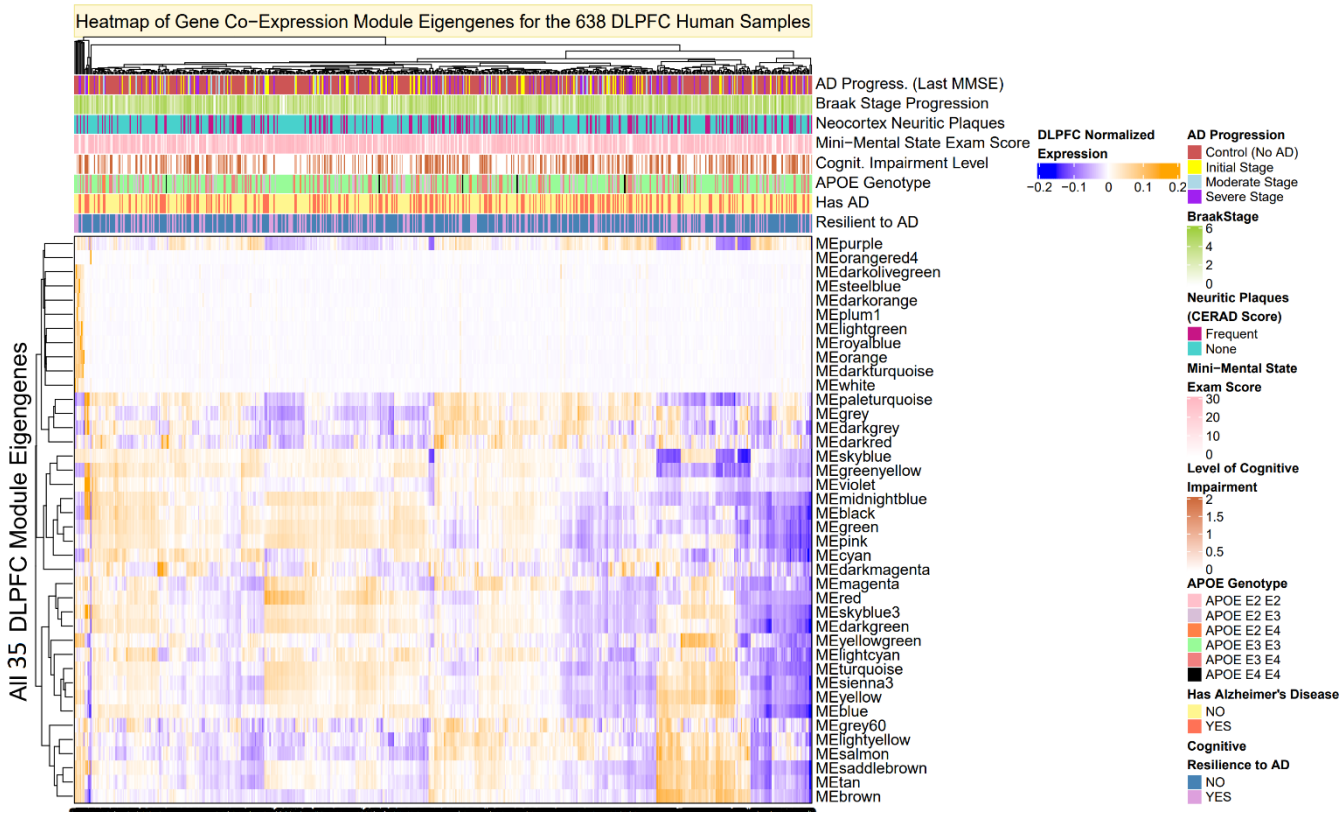


Figure A.1E) – Heatmap of all 35 Gene Co-expression Module Eigengenes for the 638 DLPFC Human Samples.

This heatmap presents the different expression dynamic patterns during AD progression for all 35 gene modules in the DLPFC region after Weighted Gene Co-Expression Network Analysis with k-means. This heatmap illustrates how the represented modular expression patterns (module eigengenes) vary for AD phenotypes, such as AD Progression (based on the last MMSE score), Braak Stage, Neo-cortex Neuritic Plaques (CERAD Score), Mini-Mental State Exam (MMSE) Score, Level of Cognitive Impairment, APOE Genotype, AD (Alzheimer's Disease), Cognitive Resilience to AD.

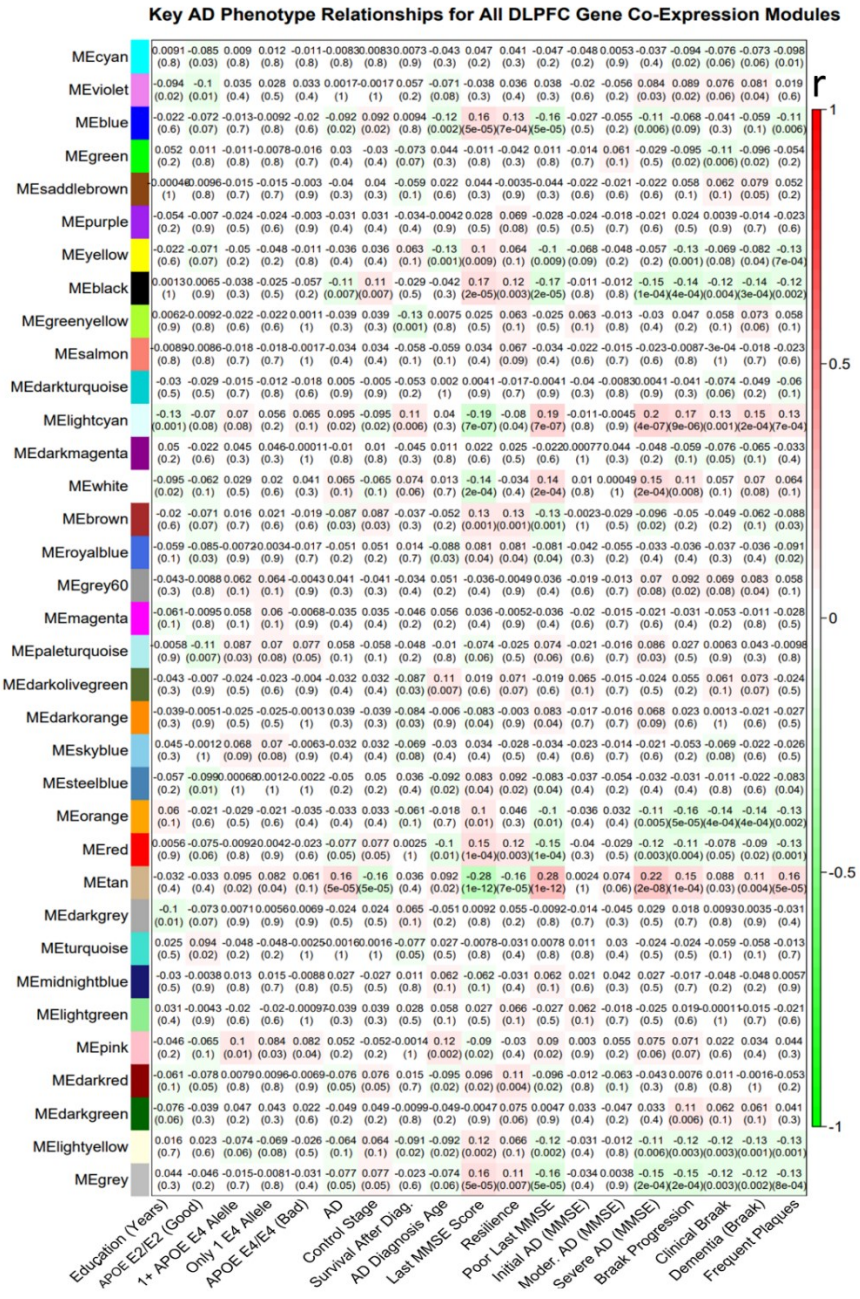


Figure A.1F) – AD Phenotype relationships for all DLPFC Gene Co-expression modules

This figure further links all 35 gene modules in the DLPFC to select respective AD phenotypes; these heatmaps present the Pearson correlation (r) on top and the resulting p-value below, for each module-phenotype relationship. These phenotypes are Education (Years), APOE E2/E2 (Good), 1+ APOE E4 Allele, Only 1 E4 Allele, APOE E4/E4 (Bad), AD, Control Stage, Survival After AD Diagnosis, AD Diagnosis Age, Last Mini-Mental State Exam (MMSE) Score, Resilience, Poor Last MMSE, Initial AD (Last MMSE), Moderate AD (Last MMSE), Severe AD (Last MMSE), Braak Progression, Clinical Braak Stage, Dementia (Braak), Frequent Plaques.

Figure A.2 Core gene Enrichments for select gene co-expression modules across the 3 brain regions: Hippocampus, LTL, and DLPFC.

These results illustrate how gene co-expression modules can be enriched for various biological functions and pathways, and how these biological functions/pathways correlate with various phenotypes. Overall, these results may help shed light on the roles of non-coding SNPs in cellular outcomes and pathways.

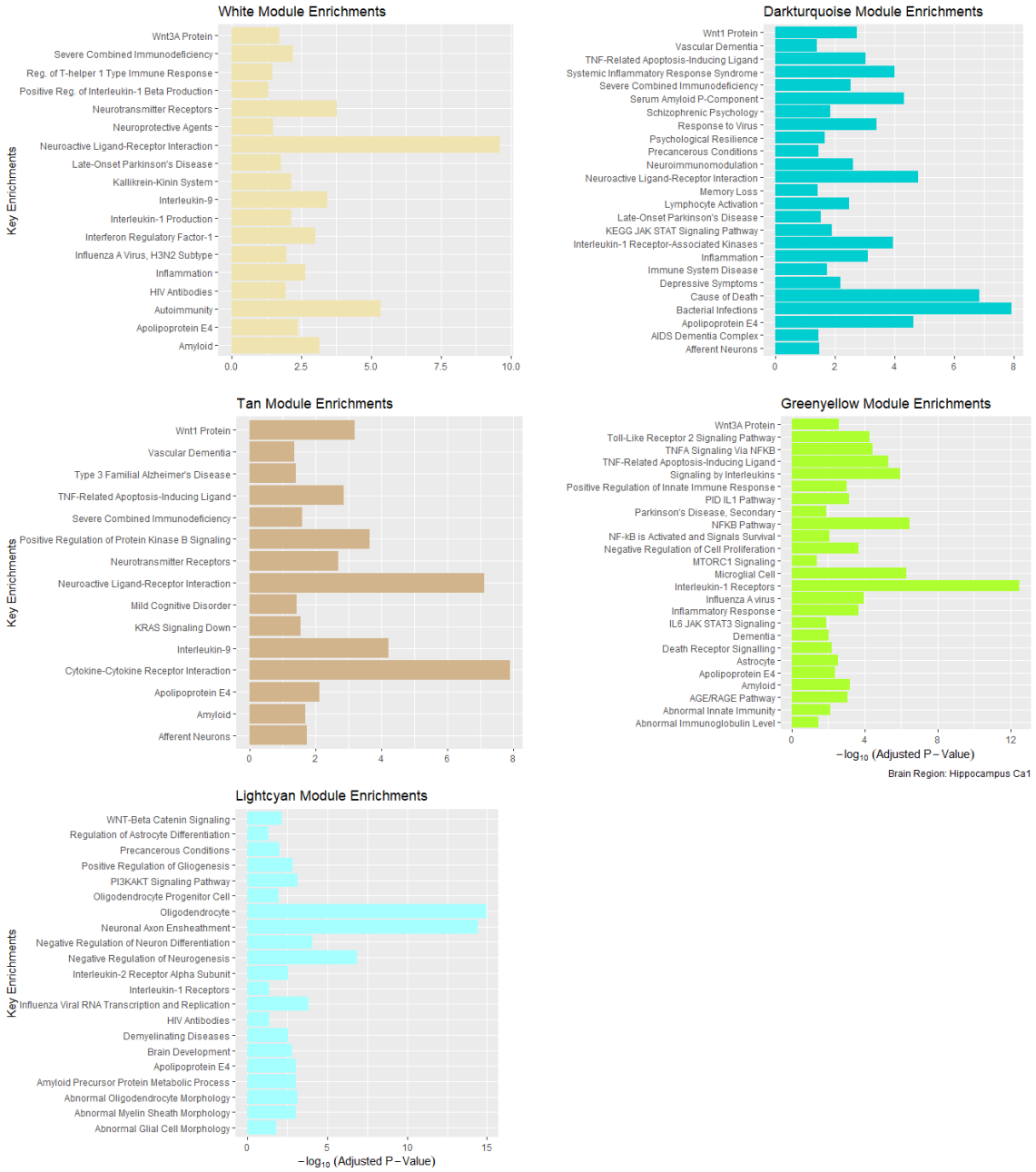


Figure A.2A) – 5 Select Gene Co-Expression Module Enrichments for the Hippocampus Ca1.

This figure illustrates 5 of the 30 gene co-expression modules for the Hippocampus Ca1 region and some of the biologically meaningful enrichments for them (carefully culled from a variety of different sources as mentioned in Table A.1). Here, the white, darkturquoise, tan, greenyellow, and lightcyan modules were found

to have strong biological significance as well as strong and significant correlations for AD phenotypes: white, darkturquoise, tan, and greenyellow modules are significantly correlated with the Severe AD Stage. Tan and greenyellow modules are also correlated with AD progression. The lightcyan module is correlated with AD progression along with Control Stage and other phenotypes.

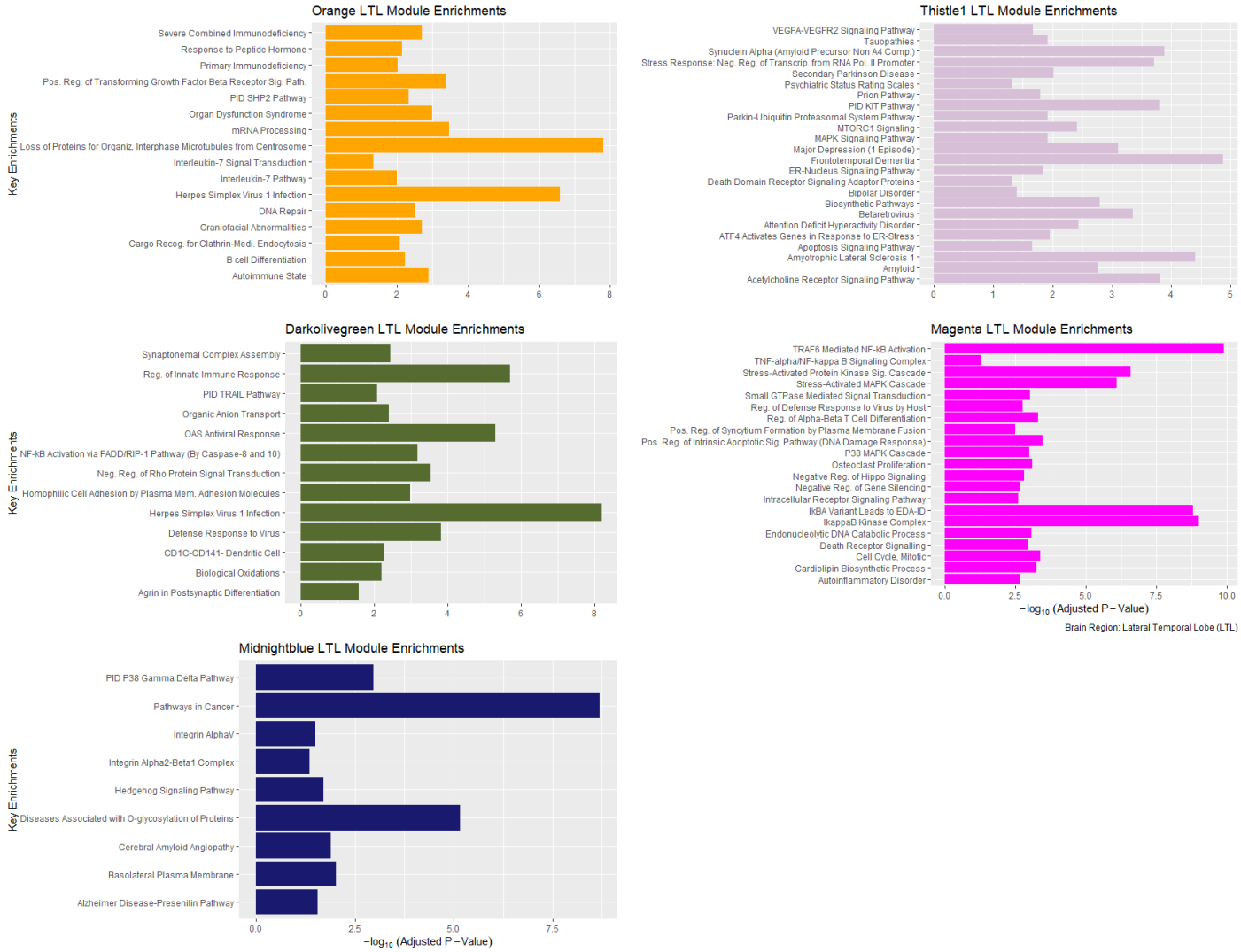


Figure A.2B) – 5 Select Gene Co-Expression Module Enrichments for the Lateral Temporal Lobe.

This figure illustrates 5 of the 56 gene co-expression modules for the LTL region and some of the biologically meaningful enrichments for them (carefully culled from a variety of different sources as mentioned in Table S1). Here, the orange, thistle1, darkolivegreen, magenta, and midnightblue modules were found to have strong biological significance as well as strong and significant correlations for AD phenotypes: orange, thistle1, darkolivegreen, and magenta gene co-expression modules are significantly correlated with AD and dementia-related phenotypes. The midnightblue module is positively correlated with aging and clinical Braak stage.

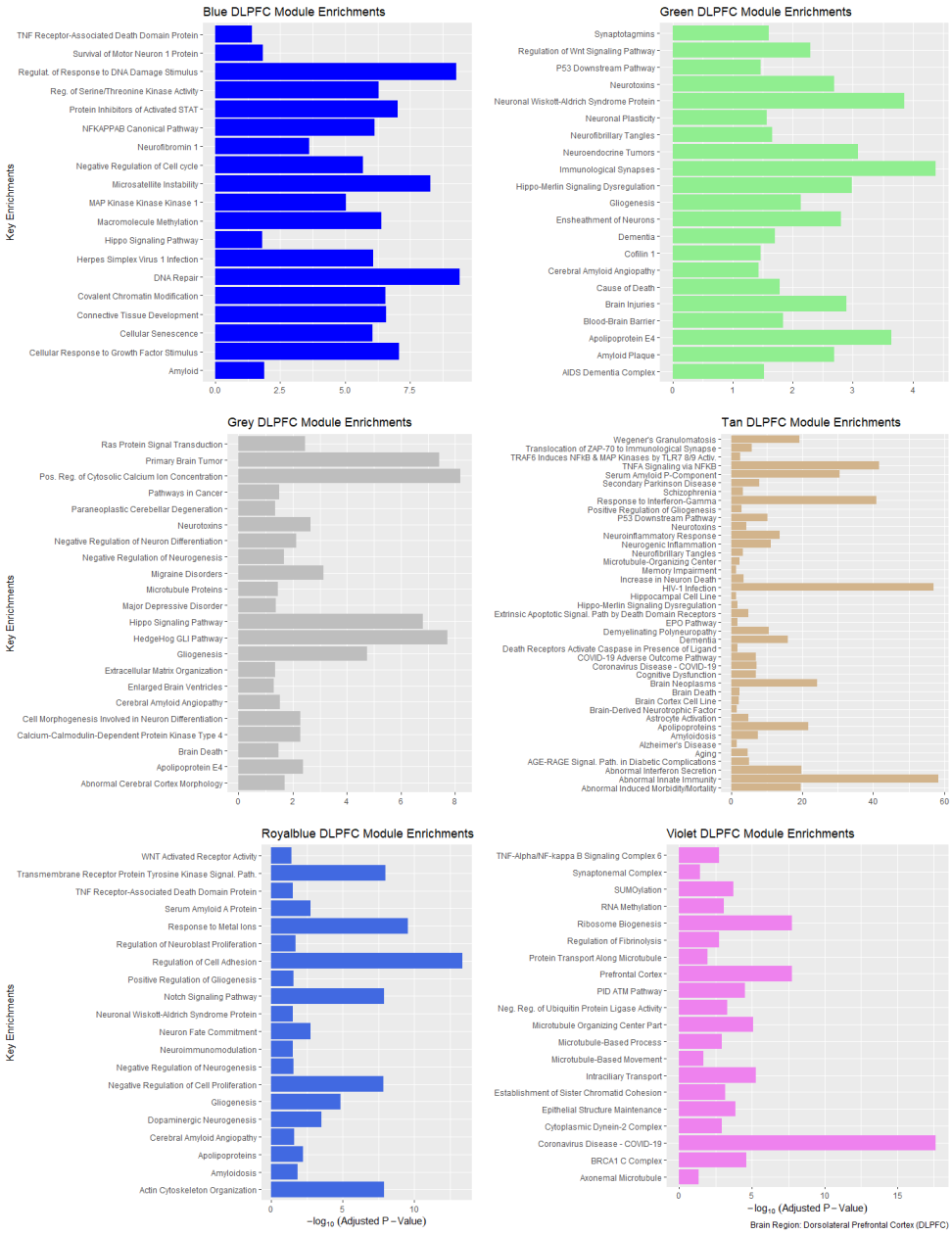


Figure A.2C) – 6 Select Gene Co-Expression Module Enrichments for the Dorsolateral Prefrontal Cortex (DLPFC) region.

This figure illustrates 6 of the 35 gene co-expression modules for the DLPFC region and some of the biologically meaningful enrichments for them (carefully culled from a variety of different sources as mentioned in Table S1). Here, the blue, green, grey, tan, royalblue, and violet gene co-expression modules were found to have strong biological significance as well as strong and significant correlations for AD phenotypes: the blue, green, grey, and royalblue gene co-expression modules are positively correlated with severe stage AD (based on last Mini-Mental State Examination (MMSE) score), severe cognitive impairment, Braak progression, poor MMSE score, and more. The violet module is positively and significantly associated with aging and Braak progression, along with other phenotypes. The tan module is associated with having both APOE4 alleles (so higher risk of LOAD), dementia based on last MMSE score, and aging, to name a few.

Figure A.3 Select enriched functions of gene co-expression modules for various AD phenotypes.

This figure focuses on groups of gene modules correlated with various phenotypes in the population. It uses all enrichments for individual modules for a phenotype. That is, to visualize enriched terms for a phenotype in a brain region, we averaged non-zero $-\log_{10}(adjust\ \mathrm{P})$ values for only gene modules significantly positively correlated (Pearson $r > 0$, $P < 0.05$) with that phenotype. Columns: AD phenotypes. The blue heatmap colors correspond to $-\log_{10}(adjust\ p - value)$, where darker shades correspond to more statistically significant enrichment values. The light-yellow color means the statistically significant gene modules associated with that phenotype are not enriched for that given enrichment term (across all possible biological resources used).

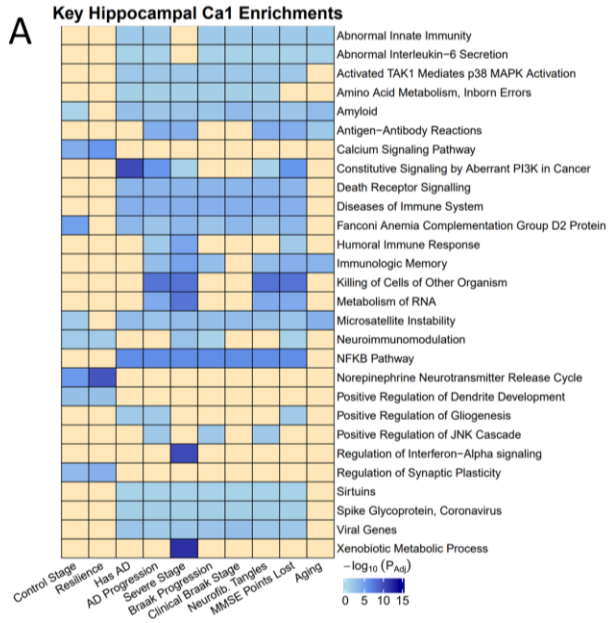


Figure A.3A) - Select enriched functions and pathways of gene modules for various AD phenotypes across the Hippocampus CA1 brain region.

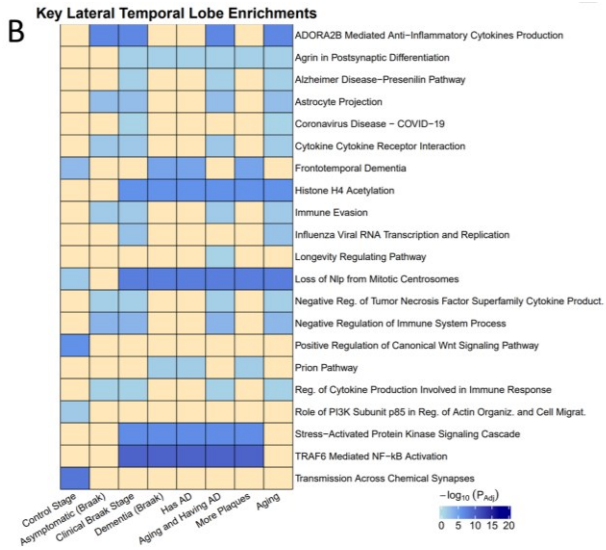


Figure A.3B) - Select enriched functions and pathways of gene modules for various AD phenotypes across the Lateral Temporal Lobe (LTL) brain region.

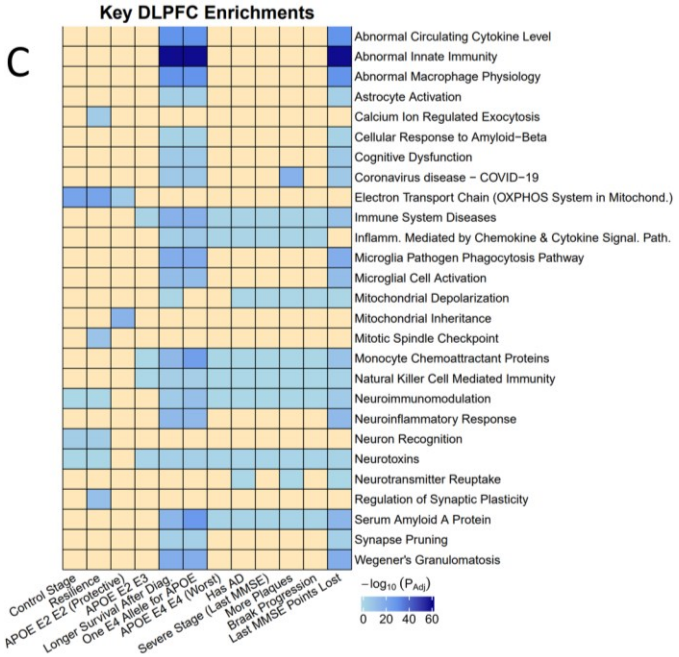


Figure A.3C) - Select enriched functions and pathways of gene co-expression modules for various AD phenotypes across the Dorsolateral Prefrontal Cortex (DLPFC) brain region.

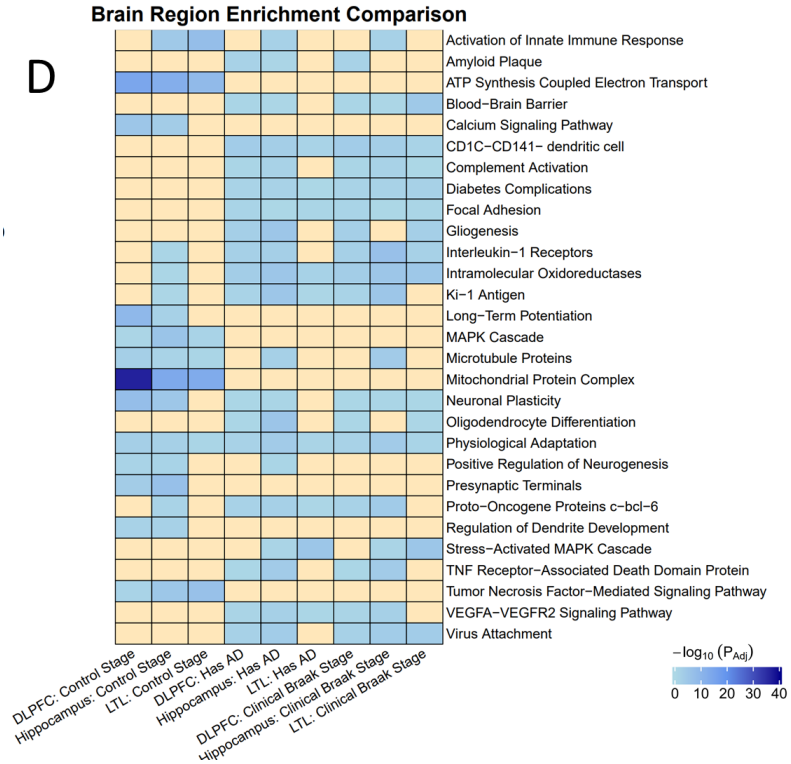


Figure A.3D) - Select enriched functions and pathways of gene co-expression modules for various AD phenotypes across the 3 brain regions (Hippocampus, Lateral Temporal Lobe (LTL), Dorsolateral Prefrontal Cortex (DLPFC)).

Figure A.4 Transcription Factor (TF) to Significantly regulated Gene module relationships.

TFs are statistically associated with gene co-expression modules in the Hippocampus and LTL based on gene expression relationships (explained in section **A.1.2**).

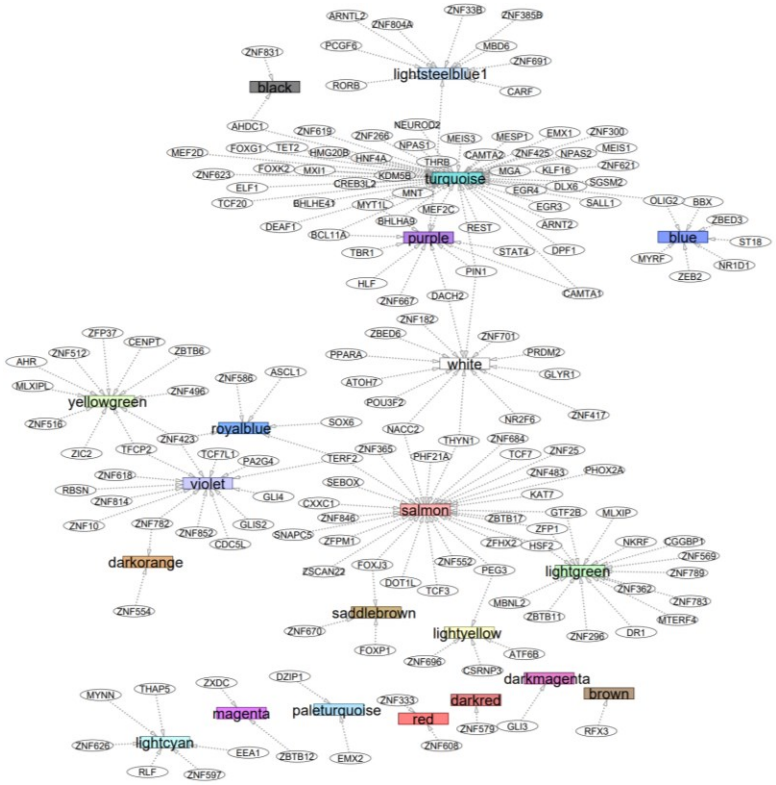


Figure A.4A) – Lateral Temporal Lobe (LTL) to Regulated Gene Module Relationships.

Circles represent TFs and rectangles represent gene co-expression modules in the Lateral Temporal Lobe (and are colored based on the module color name provided by WGCNA). Edges are directed and go from TF to Phenotype. Please note this is the network based on RTN Master Regulatory Analysis (MRA) that reveals statistically significant regulatory relationships between TFs and gene co-expression modules in the LTL. For each gene module, MRA performed enrichment analysis using the inferred GRN, the phenotype (Module Membership correlation of all genes to that module), and hits (genes assigned to that module). TFs significantly regulate 21 out of the 56 gene co-expression modules in the Lateral Temporal Lobe.

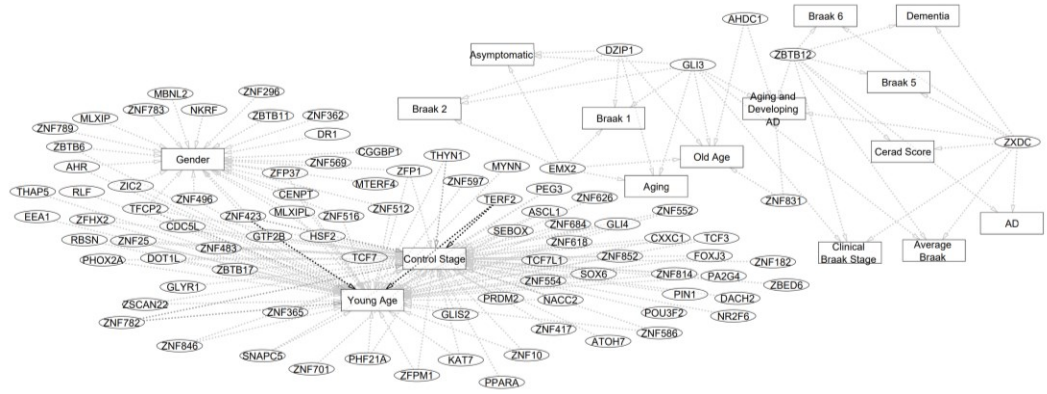


Figure A.4B) – TF to Phenotypes in Regulated Gene Modules in the Lateral Temporal Lobe (LTL). Circles represent TFs and rectangles represent phenotypes in the LTL. Edges are directed and go from TF to Phenotype. Based on **Figure A.4A** (TF to regulated gene modules in LTL), we analyzed phenotypes that are significantly positively correlated ($p < 0.05$, $r > 0$) with each respective gene module. Then, for each TF, we counted the number of gene modules with a particular phenotype that it regulates. We found that a TF would regulate between 1 and 3 modules with a particular phenotype. The greater the number of modules with that phenotype that the TF regulates, the darker the directed edge would be between them. For instance, TERF2 has the darkest edge pointing to the Control Stage Node since it regulates 3 different modules that are positively and significantly associated with the Control Stage.

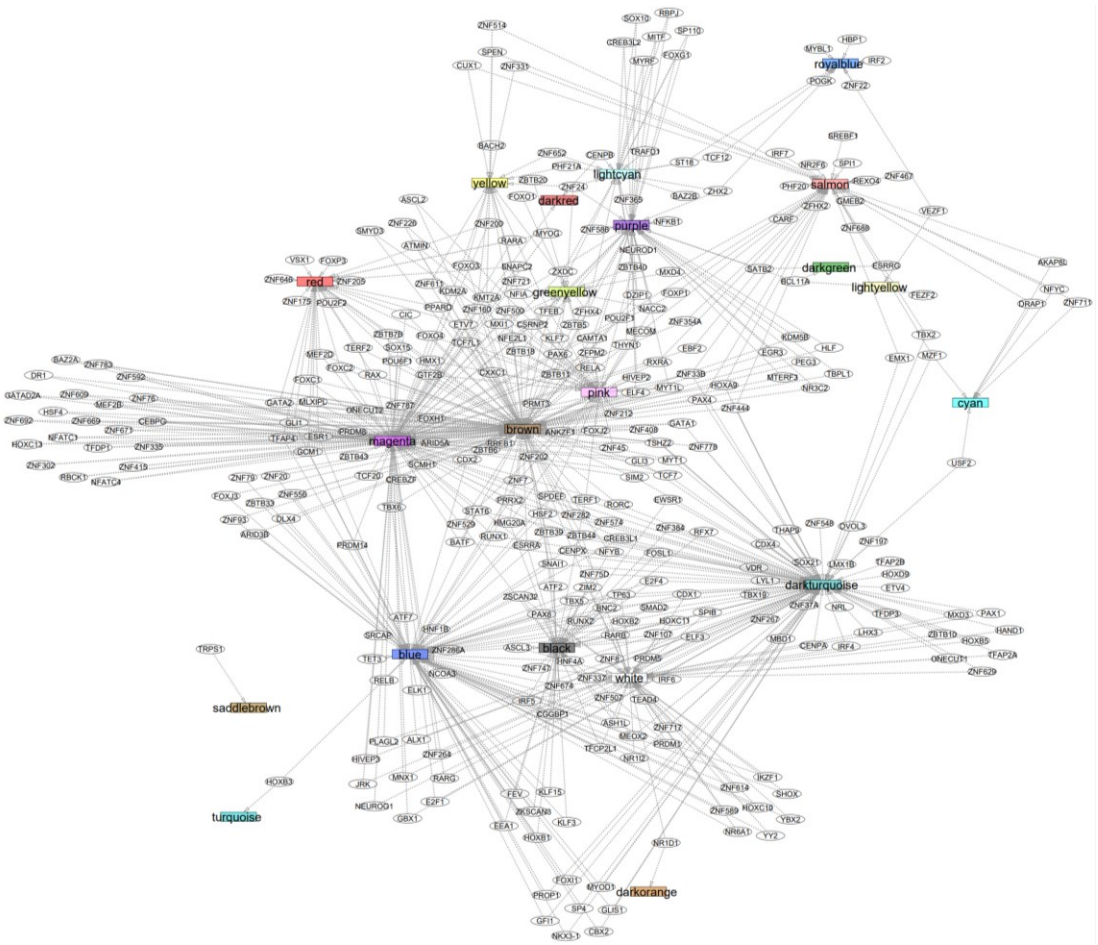


Figure A.4C) – Hippocampal TF to Regulated Gene Module Relationships to Phenotypes in Regulated Gene Modules in the Lateral Temporal Lobe.

Circles represent TFs and rectangles represent gene co-expression modules in the Hippocampus (and are also colored appropriately based on the module color name provided by WGCNA). Edges are directed and go from TF to Phenotype. Please note that this is the network based on RTN MRA that reveals statistically significant regulatory relationships between TFs and the gene co-expression modules in the Hippocampus. We found TFs significantly regulate 21 out of the 30 gene co-expression modules in the Hippocampus.

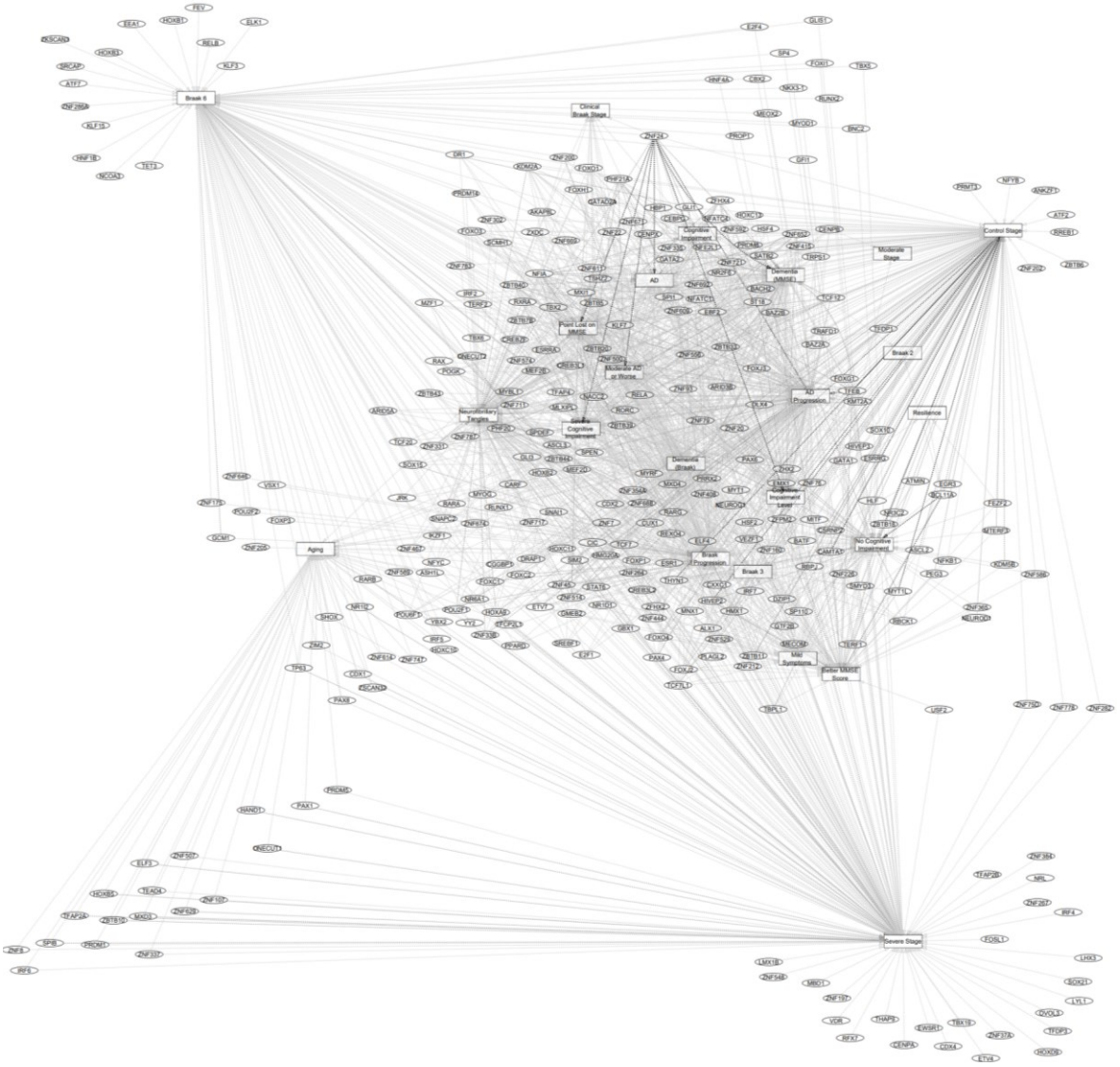
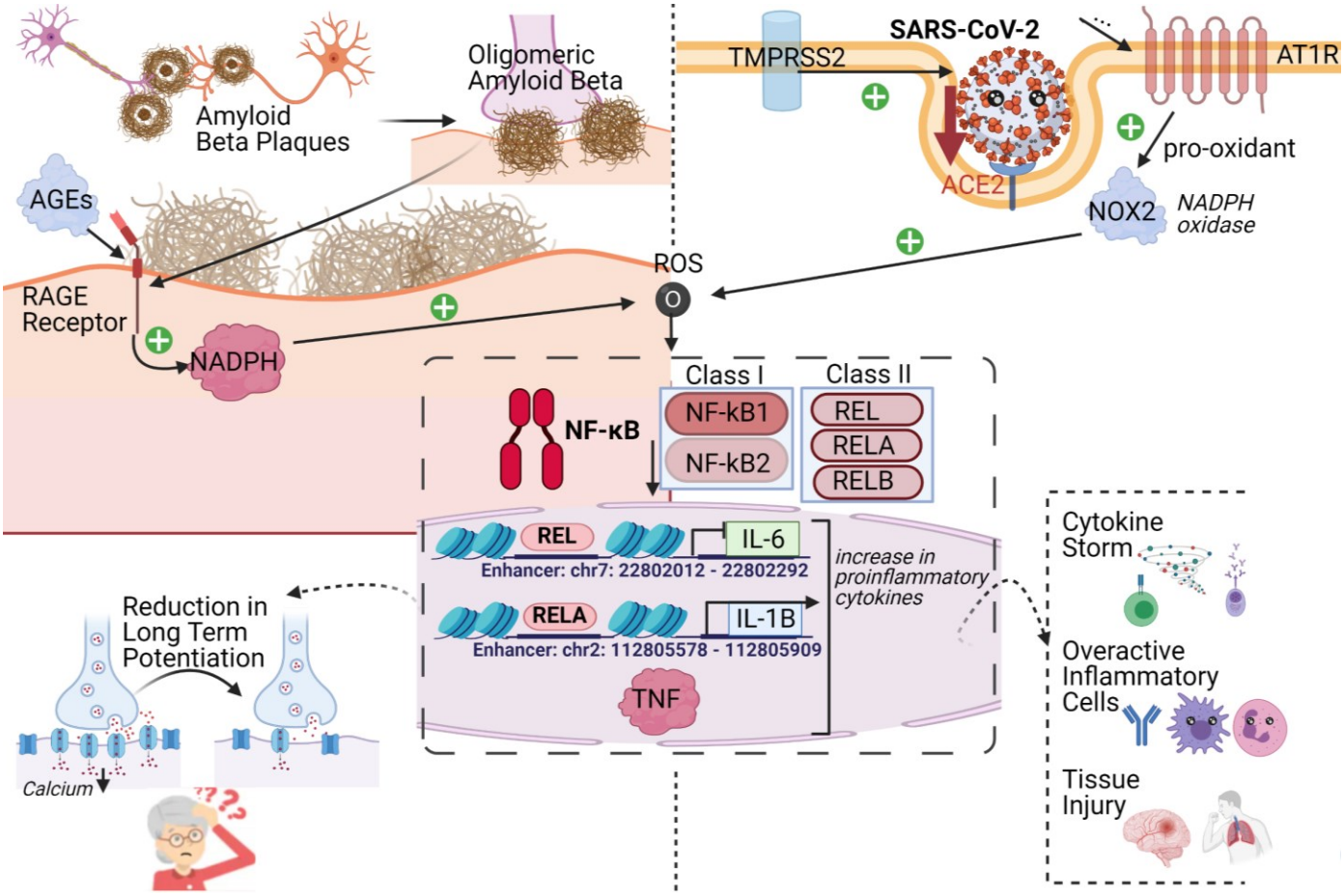


Figure A.4D) – TF to Phenotypes in Regulated Gene Module Relationships in the Hippocampus.
Circles represent Transcription Factors (TFs) and rectangles represent phenotypes in the Hippocampus. Edges are directed and go from TF to Phenotype. Based on the Network in **Figure A.4C** (TF to the regulated gene modules in the Hippocampus), we analyzed the phenotypes that are significantly positively correlated ($p < 0.05$, $r > 0$) with each respective gene module. Then, for each TF, we counted the number of gene modules with a particular phenotype that it regulates. We found that a TF would regulate between 1 and 3 modules with a particular phenotype. The greater the number of modules with that phenotype that the TF regulates, the darker the directed edge would be between them. For instance, TFEB, THYN1, ZBTB18, CSRNP2, CAMTA1, HIVEP2, ZFPM2, MYT1L, and PAX6 have the darkest edge pointing to the Control Stage Node since they regulate 3 different modules that are positively and significantly associated with the Control Stage. On the other hand, ZNF24 has the darkest edge pointing to the AD Progression, AD, Severe Cognitive Impairment and Moderate Stage or Worse nodes as it regulates 3 different modules positively associated with those respective phenotypes. Interestingly, TFEB regulates both 3 AD progression modules and 3 control stage modules.

Figure A.5 Shared Covid-19 and Alzheimer's disease (AD) Pathways: Correlations with AD) in the Hippocampus CA1 brain region.



Please note that this visualization presents a simplified version of AD (left hand side of vertical dashed line in middle of diagram) and Covid-19 (right hand side) and common mechanisms, based on the KEGG pathways (from Pathview) and findings from our TF-TG Hippocampus GRN Network. Mechanisms in red, such as NF-kB, tend to be increased in expression in the AD Hippocampus. This NF-kB pathway is associated with highly-conserved TFs that are involved in various cellular processes including regulation of inflammation, cell growth, and apoptosis SARS-CoV-2 results in excess NF-kB activation by either TLR4-mediated NF-kB activation or ER stress-induced NF-kB activation NF-kB is composed of 5 different TFs or NF-kB signaling pathway genes, which belong to 2 different classes: Class 1 (NF-kB1 and NF-kB2) and Class 2 (REL, RELA, and RELB). Individuals with AD usually have an increased expression of the RAGE receptor and NADPH that impacts AD outcomes. They also have a greater expression of AT1R and ACE2 receptors, which may be associated with exacerbated Covid-19 outcomes. In AD and Covid-19, 2 NFKB TFs (RELA and NFKB1), are activated and enter the nucleus to help transcribe various pro-inflammatory proteins including cytokines. Here, we focus on cytokines, which are secreted by certain immune cells and have an impact on the gene expression of target cells; in both AD and Covid-19, NF-kB regulation of IL-6, IL-1B, and TNF is involved in adverse events. In fact, soluble TNF kills cells infected with viruses/bacteria but in excess may trigger neurodegeneration, inflammation, neuronal death, and destruction of healthy tissues.

Figure A.6 Correlations of the NFKB TFs with Various AD Phenotypes across brain regions.

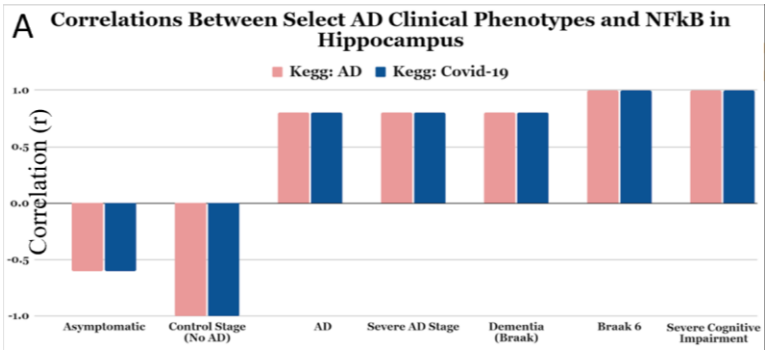


Figure A.6A) - Correlations of NFKB pathway (KEGG: hsa05171) and AD pathway (KEGG: hsa05010) with AD phenotypes from the Pathview analysis of Hippocampal expression data of pathway genes. Human KEGG pathways (homo sapiens).

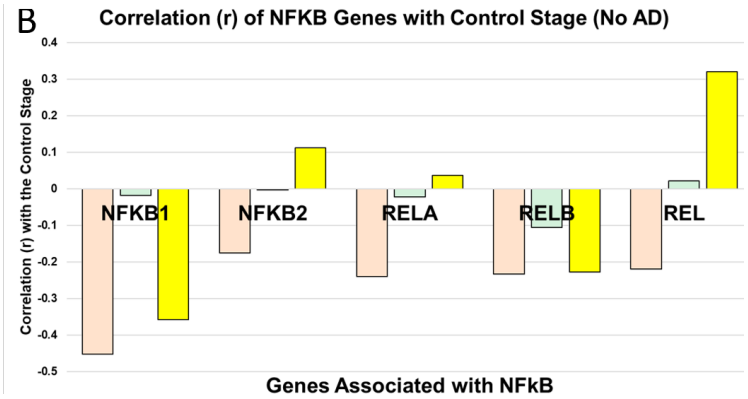


Figure A.6B) - Correlations of NFKB TFs (NFKB1, NFKB2, RELA, RELB, REL) with Control in 3 regions.

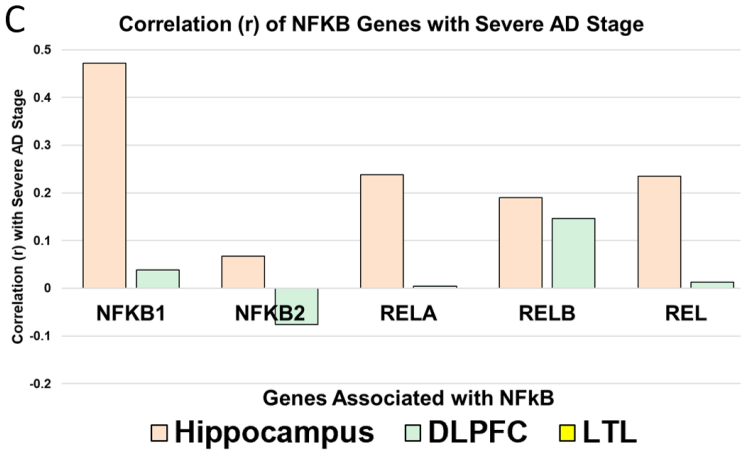


Figure A.6C) - Correlations of NFKB TFs with Severe stage in Hippocampus and DLPFC.

Figure A.7 Regulation of Additional Covid-19 Cytokines by NFKB TFs**.**

Please note that this Figure expands on the Cytokines in **Figure 2.3C** and explores how they are regulated by TFs that themselves are regulated by NFKB TFs (RELA and/or NFKB1). These cytokines are involved in the cytokine storm associated with Covid-19: **(A)** *IL-1B*, **(B)** *MMP1*, **(C)** *CCL2*, and **(D)** *MMP3*.

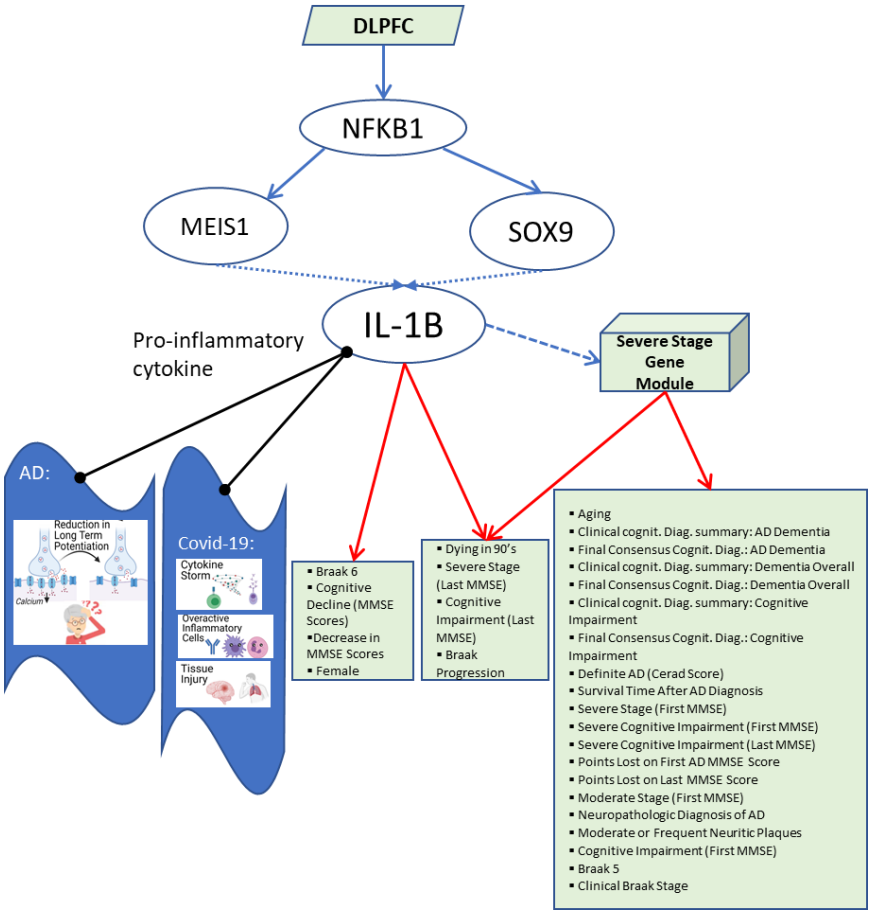


Figure A.7A) – Regulation of Cytokine *IL-1B* in the DLPFC.

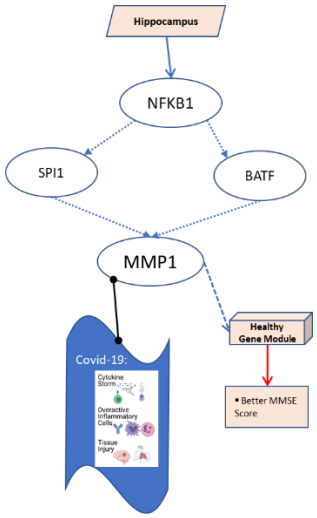


Figure A.7B) – Regulation of cytokine *MMP1* in the Hippocampus.

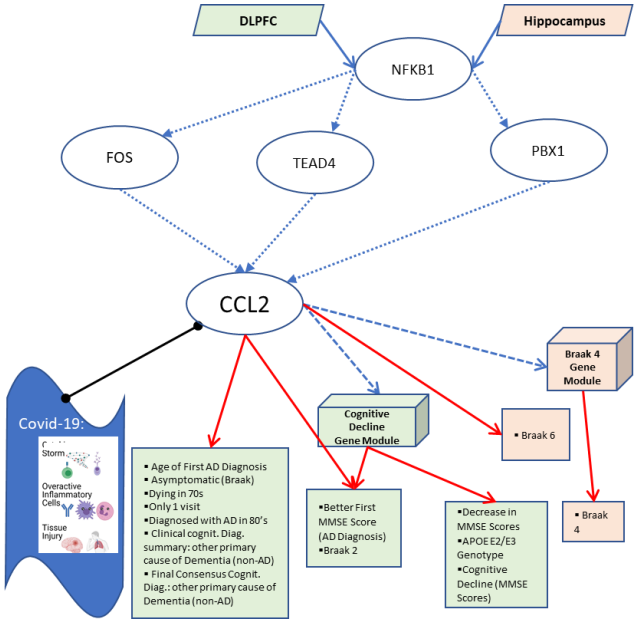


Figure A.7C) – Regulation of cytokine *CCL2* in the Hippocampus and the DLPFC.

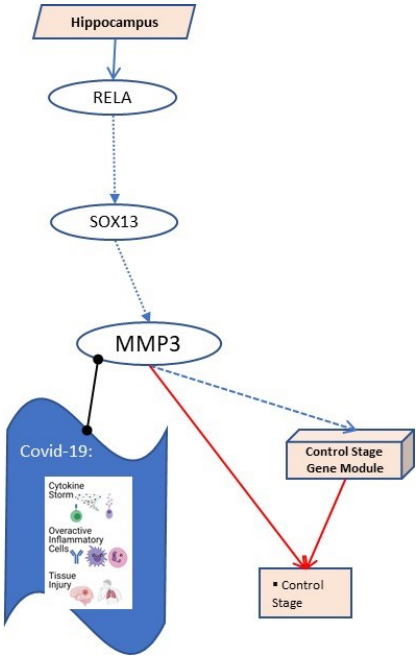
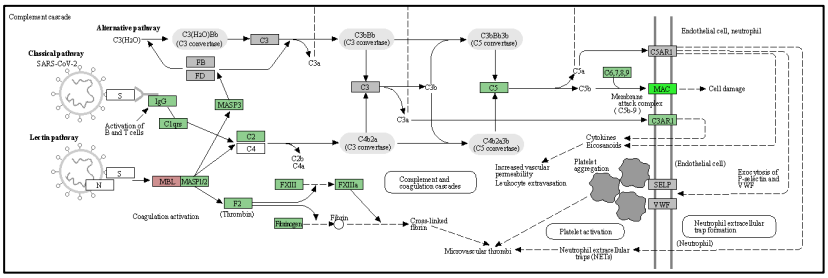


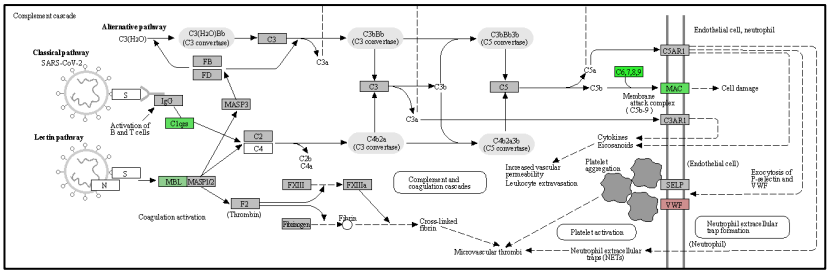
Figure A.7D) – Regulation of cytokine *MMP3* in the Hippocampus.

Figure A.8 Correlation of Various Hippocampal AD Stages with Covid-19 Complement Cascade KEGG Pathway (by Pathview):

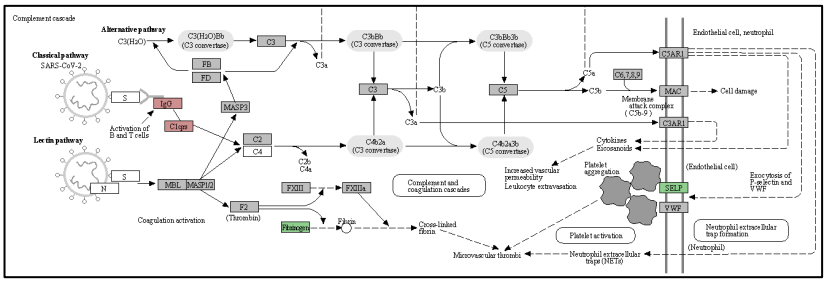
Control Stage (No AD)



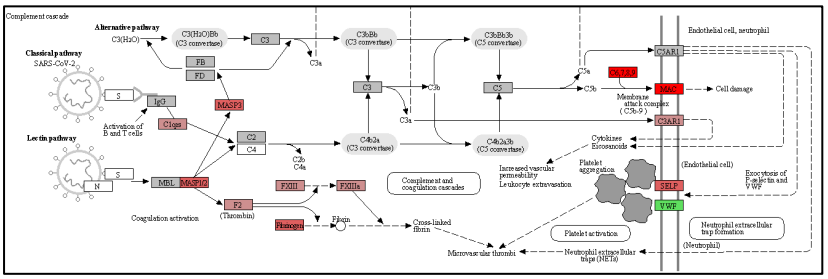
Initial Stage of AD



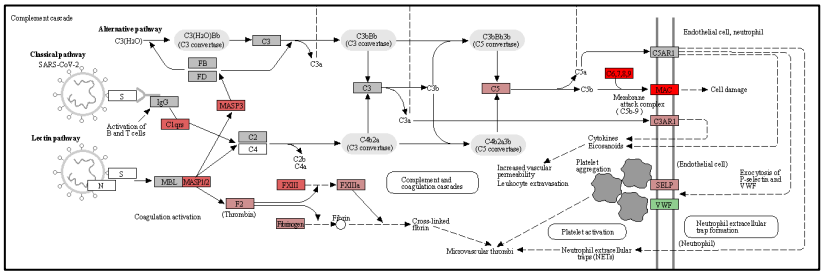
Moderate Stage of AD



Severe Stage of AD



AD Progression



This illustrates how AD progression is associated with a greater expression (higher correlation r, red) of the Complement Cascade Components of the immune system in the Coronavirus Disease. As AD progresses from the Initial to Moderate to Severe Stages, many more of the mechanisms are colored red, which indicates that those mechanisms tend to be correlated positively ($r > 0$) with AD progression. Mechanisms colored in green tend to be negatively correlated ($r < 0$) with AD progression.

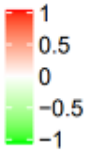
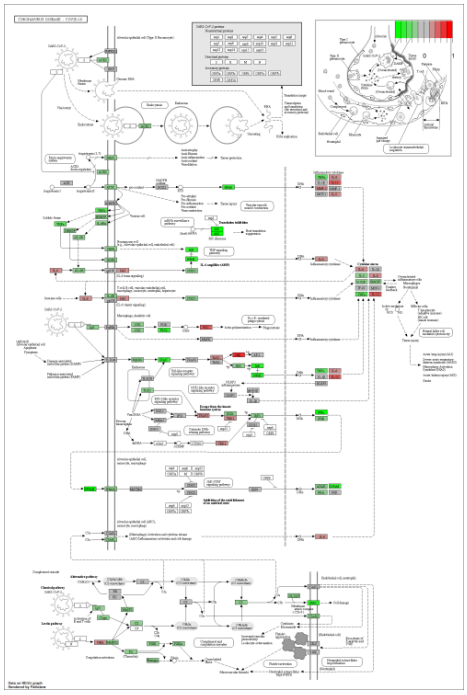
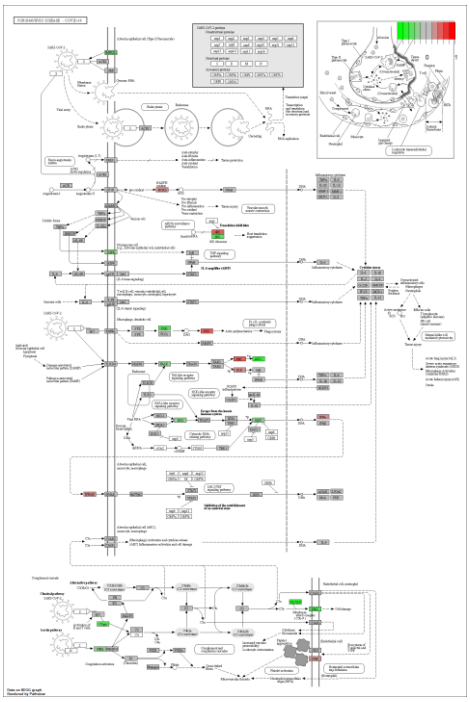


Figure A.9 Hippocampal CA1 Brain Region: Correlation of Various AD stages with Coronavirus Disease (Covid-19) KEGG Pathway. This figure zooms out on hsa05171 and shows how AD stages in the Hippocampus are correlated with various mechanisms in Covid-19.

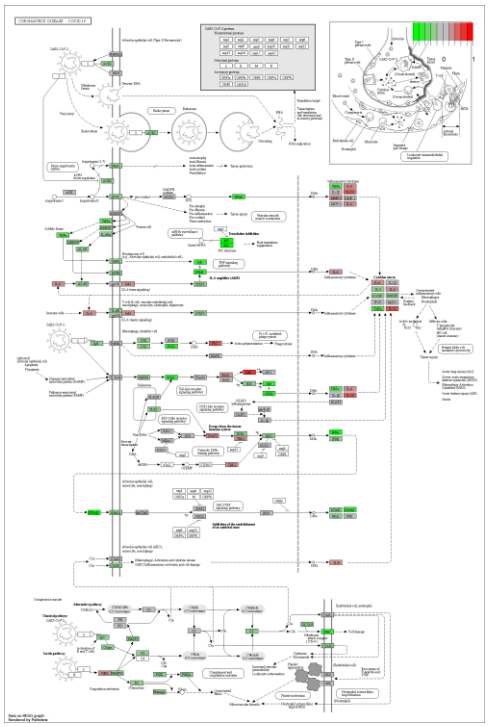
Control Stage (No AD)



Initial Stage of AD



Moderate Stage of AD



Severe Stage of AD

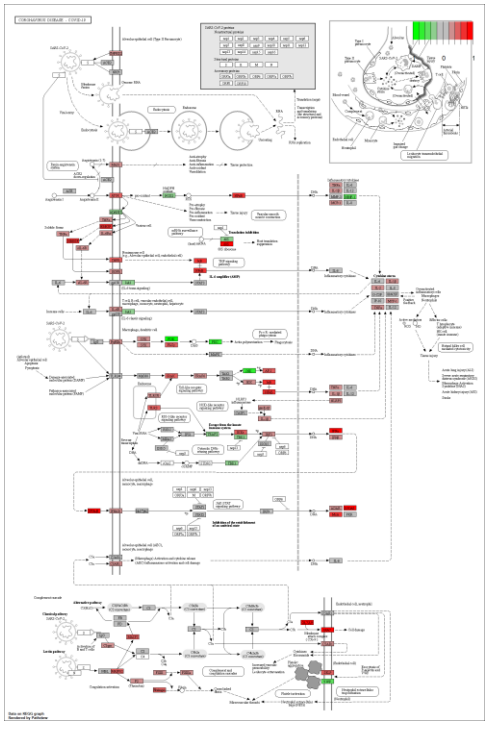
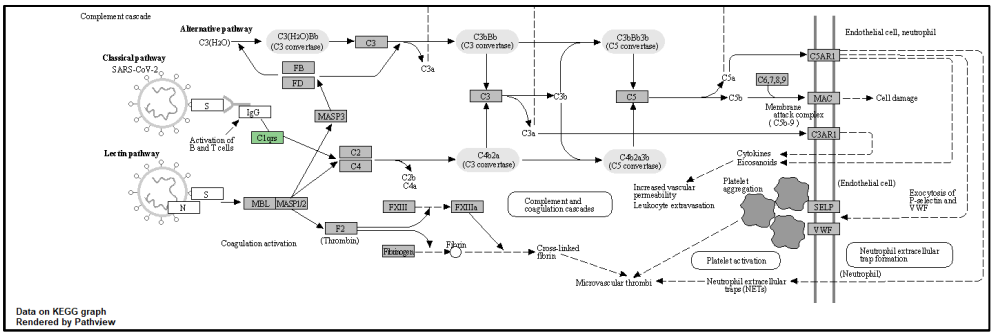
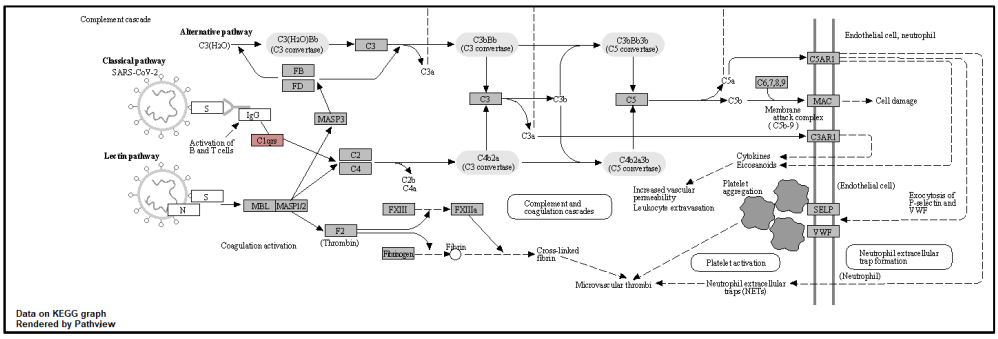


Figure A.10 DLPFC Brain Region: Correlation of Various APOE Genotypes with the Complement Cascade in the Coronavirus Disease (Covid-19) KEGG Pathview Pathway.

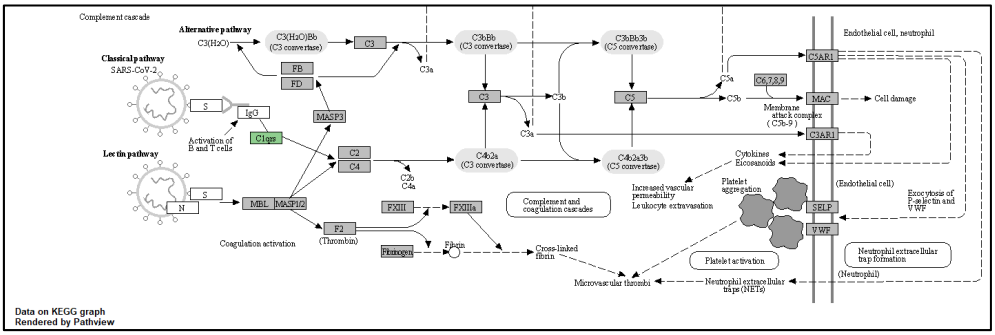
APOE E2/E2 (Protective Against AD)



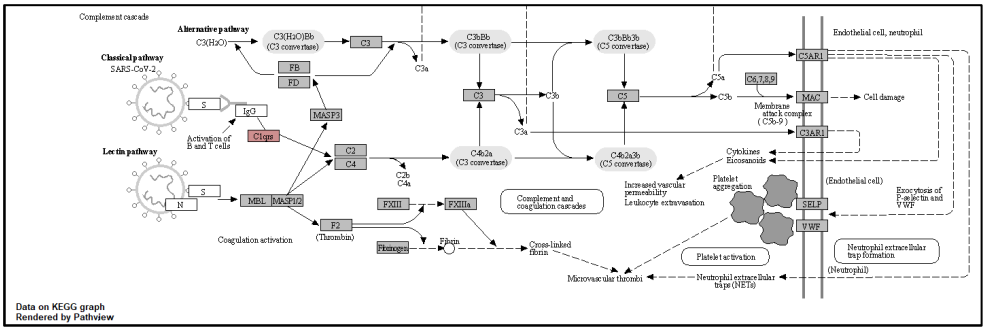
Exactly 1 APOE E4 Allele (Increased Risk of AD): APOE E2/E4 or APOE E3/E4



APOE E3/E3 (Average Risk of AD)



APOE E4/E4 (Highest Risk of AD and Higher Risk of Severe Covid-19 Outcomes)



Pathview (Luo and Brouwer 2013a) is used to obtain the KEGG pathways and color them based on correlations with APOE genotype. In the Complement Cascade of the Coronavirus Pathway, we observe consistent differences between APOE genotype and expression of C1qrs (Complement Component 1 Complex).

Figure A.11 Correlations Between AD Phenotypes and KEGG Pathways (Covid-19 and AD) in Various Brain Regions:

This Correlation Plot has 10 main phenotype groups and analyzes differences in the Pearson correlation (r) between these phenotypes and common shared mechanisms between AD and Covid-19. Usually, a phenotype will have the same correlation in AD and Covid-19 for a shared mechanism (except in few situations where a shared mechanism between AD and Covid-19 Kegg Pathways may involve slight differences in proteins. Ex. IKK in the LTL). The mechanisms are hierarchically clustered in each brain region based on these correlations.

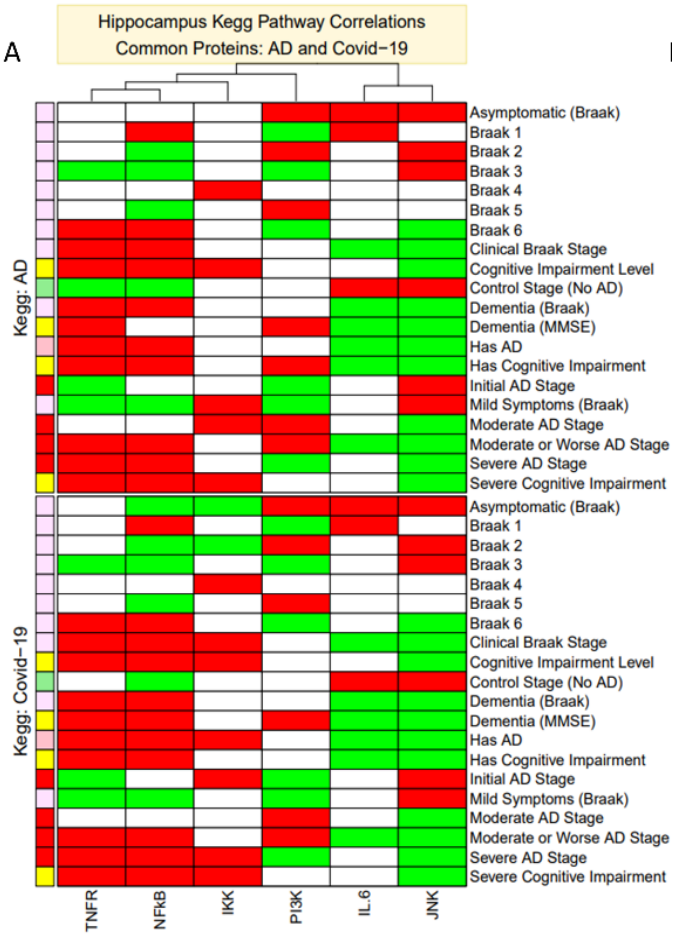


Figure A.11A) - Correlation Plot in the Hippocampus CA1.

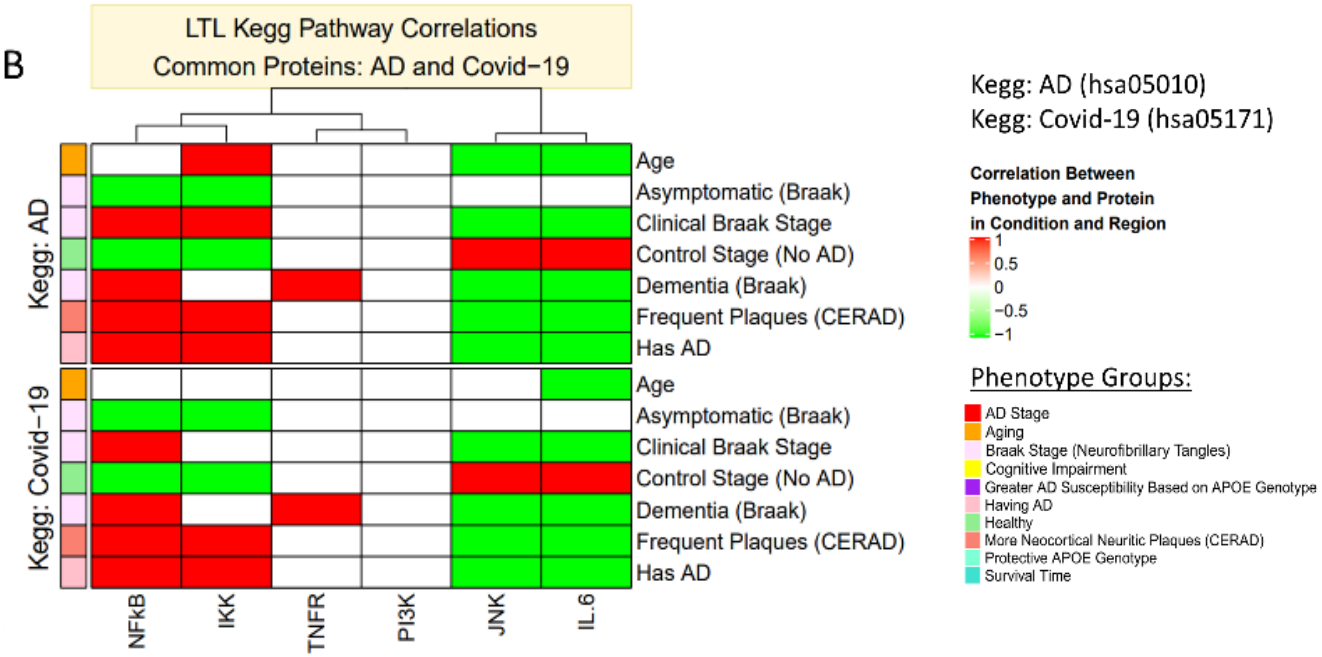


Figure A.11B) - Correlation Plot in the Lateral Temporal Lobe (LTL).

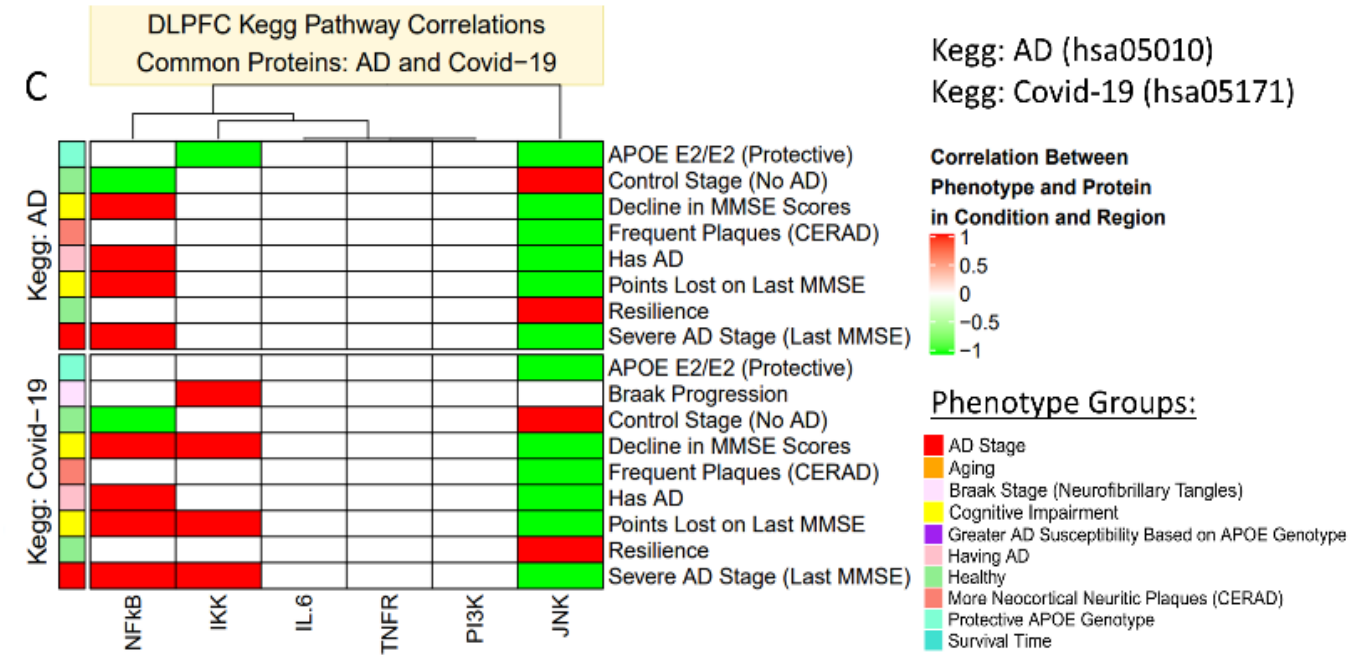


Figure A.11C) - Correlation Plot in the Dorsolateral Prefrontal Cortex (DLPFC).

Figure A.12 Comparison of Brain Regions: Correlation of Having Alzheimer's Disease (AD) with the Coronavirus Disease (Covid-19) KEGG Pathway hsa05171

This figure illustrates differences in Covid-19 related mechanisms associated with the Alzheimer's disease (AD) phenotype in the 3 brain regions.

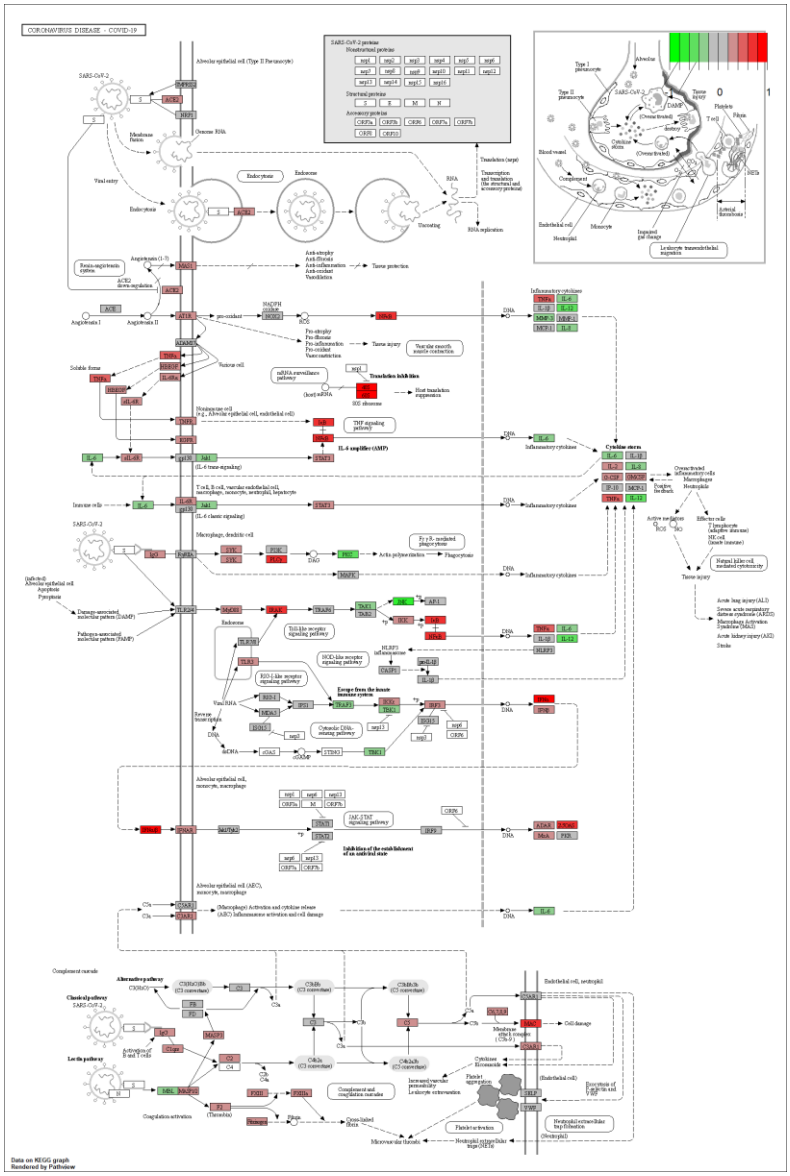


Figure A.12A) - Hippocampus CA1 brain region.

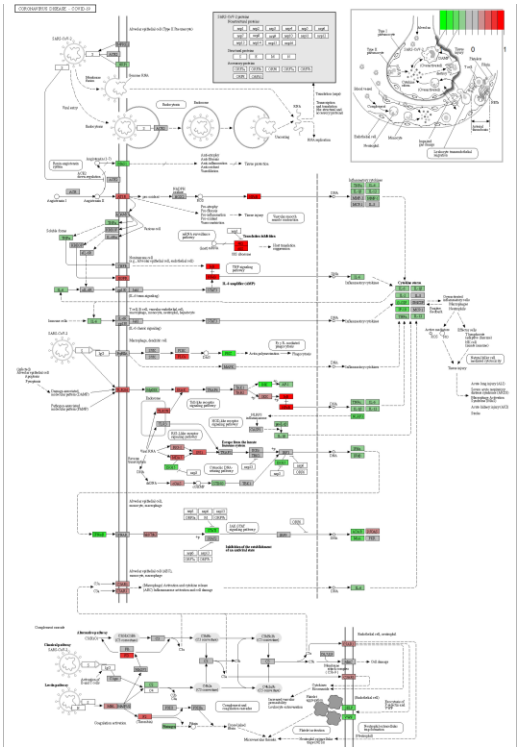


Figure A.12B) - Lateral Temporal Lobe (LTL) brain region.

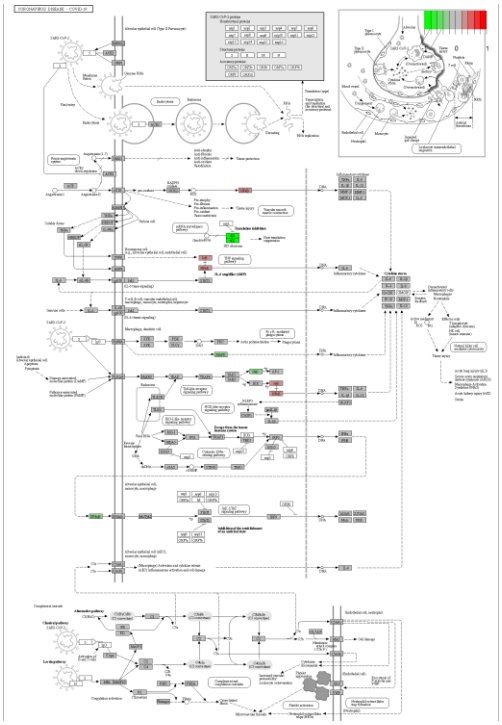
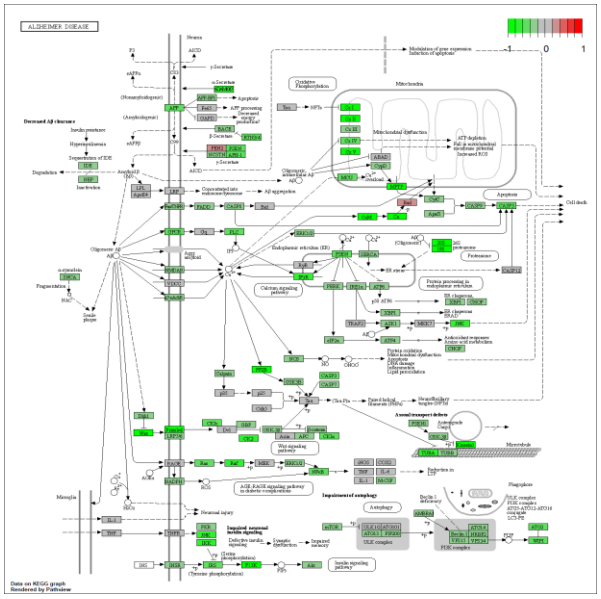


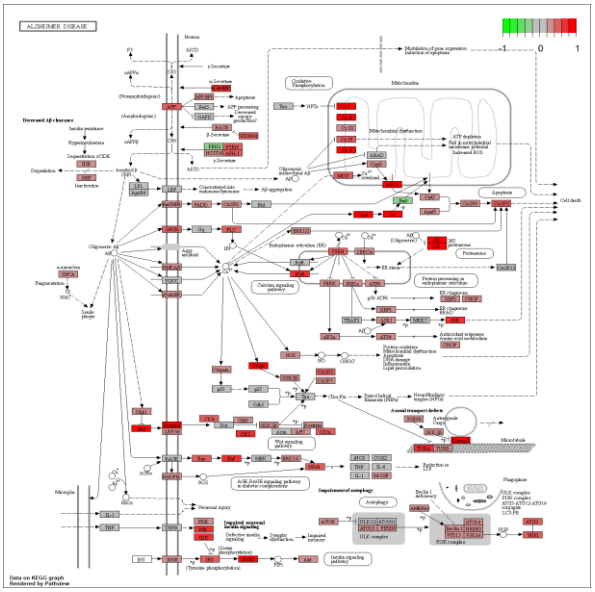
Figure A.12C) - Dorsolateral Prefrontal Cortex (DLPFC) brain region.

Figure A.13 Correlation of Having Covid-19 with the Alzheimer's Disease KEGG Pathway.

Covid-19 Negative

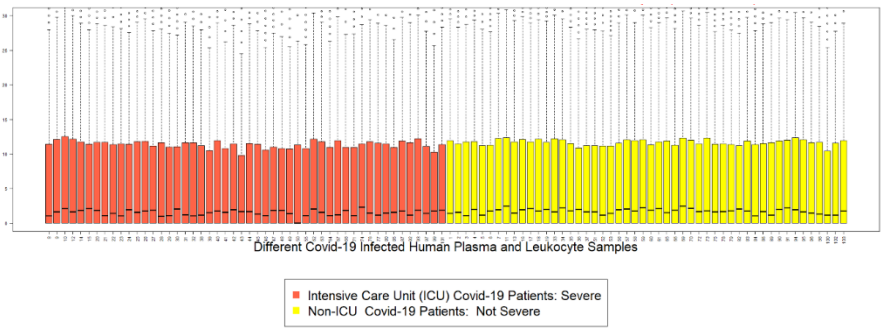


Covid-19 Positive



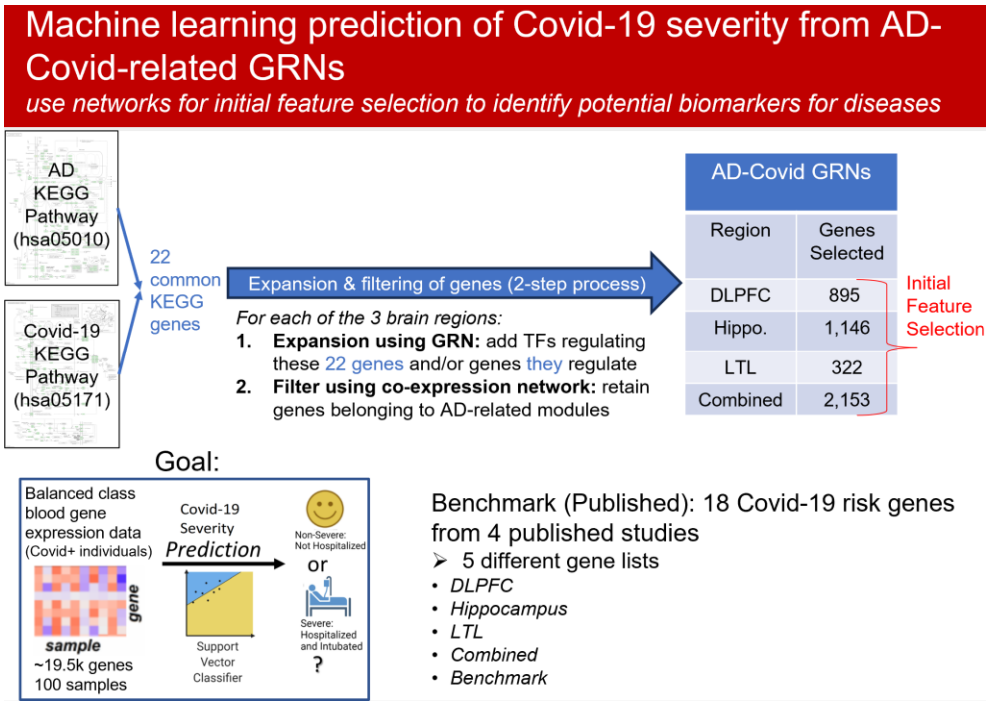
This figure uses the 126 samples from GSE157103 (or GSM4753022, where 100 are Covid-19 positive (infected with SARS-CoV-2 virus) and 26 are negative) and respective gene expression data. Covid-19+ individuals have a 1 for the "Covid-19 Positive" binary phenotype, and those who are negative have a 0 for this phenotype. Similarly, the "Covid-19 Negative" phenotype is 0 for those with Covid-19 and 1 for those without Covid-19. Correlation r was computed between all of the genes in the expression data set and these Covid-19-related phenotypes. Those resulting correlation values were then provided to Pathview (using all respective defaults) for AD KEGG pathway graph (hsa05010). Then, correlation was computed between the mechanisms in AD and each of the 2 Covid-19 phenotypes. As expected, mechanisms that are negatively associated with being Covid-19 negative (green colors) are positively associated with having Covid-19 (red colors), as the 2 phenotypes are inverse (opposites). The results illustrate that having Covid-19 (being Covid-19 positive) is associated with greater expression of many mechanisms involved in AD, which may support our work regarding finding a possible AD-Covid-19 link.

Figure A.14 Median Normalized Total Gene Read Counts for 100 Covid-19 Infected Human Samples.



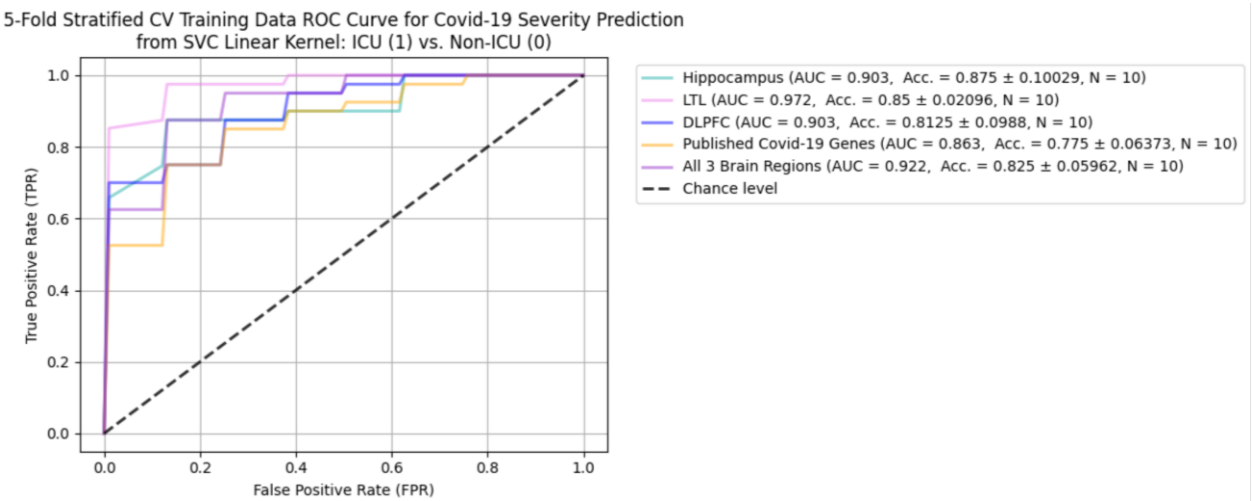
Median Normalization was performed on all genes in GSM4753022 (Overmyer et al. 2021) for the 100 Covid-19+ samples only. This Boxplot illustrates the results between the 2 groups (ICU and non-ICU). This data was used for Differential Gene Expression Analysis to find Differentially Expressed Genes (DEGs).

Figure A.15 Machine learning prediction of Covid-19 severity from AD-Covid-related GRNs.



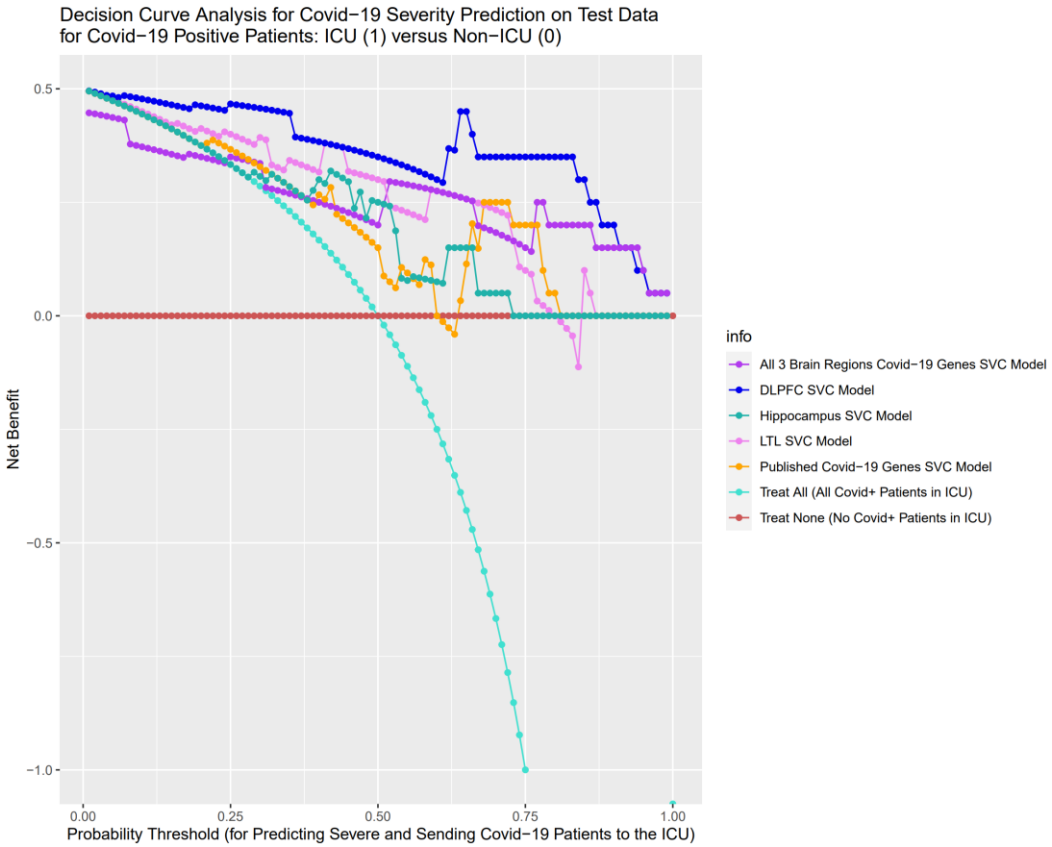
Our networks can be applied machine learning tasks to identify biomarkers for conditions. Our GRNs and our gene co-expression networks have been linked to various AD phenotypes in the population for our 3 brain regions. So, our goal is to try to use these GRNs and our gene co-expression networks to perform feature selection to identify potential biomarkers for Covid-19 severity and by extension for neuroinflammation. That is, from blood gene expression data for Covid+ individuals, we try to predict which patients will have severe outcomes (be hospitalized in the ICU intensive care unit) and which will not be (non-severe). This is balanced prediction since we have an equal number of non-severe and severe patients. So, around 19 and a half thousand genes, 50 severe patients, 50 non-severe patients. To this end, we look at the KEGG network pathways for AD and Covid-19 and identify 22 common genes. We use our GRNs to identify any genes that are directly connected to those 22: they could be TFs that regulate them and/or TGs of those 22 genes. So the expansion step blows up this list of genes. Then, we filter to keep only genes belonging to an AD-related phenotype module. 21 out of 30 gene modules in Hippocampus are AD-related (10 are positive for AD), 28 out of 56 gene modules in LTL are AD-related (7 are positive for AD), and all 36 gene modules in DLPFC are AD-related (18 are positive for AD). We do this for each brain region and then also have a combined list based on the 3 regions. Further, we pool together 18 Covid-19 genes from 4 different published studies, which we use as our benchmark. So, we have 895 initial features for our DLPFC, 1,146 for the Hippocampus, 322 for the LTL, 2,153 combined, and 18 benchmark. We fit a Support Vector Classifier model to predict Covid-19 severity based on these features. We used 80% of our data for training and 20% for testing. We perform recursive feature elimination (RFE) with 5-fold stratified cross-validation using an SVM model to select the # of optimal benchmark genes and obtain 10 top genes. Then, we run RFE to select the top 10 optimal genes for each of our 5 model. Then, each model is trained on those 10 optimal genes. For ease of comparison, we keep this # of genes fixed.

Figure A.16 Receiver Operating Characteristic (ROC) curves and area under curve (AUC) values for classifying Covid-19 severity in the 80 samples in the training data.



Please note that stratified 5-fold cross validation (CV) was conducted and standard error values were collected for each of the CV folds when calculating the AUC. Each fold has 16 unique samples held out (8 severe, 8 non-severe) and 64 samples used for training (32 severe, 32 non-severe) and each sample is held-out exactly once. Our average 5-fold stratified CV areas under the ROC curve (AUC) for training data are larger than the benchmark (0.903 for Hippocampus, 0.903 for DLPFC, 0.972 for LTL, 0.9217 for combined regions, 0.86 for benchmark). Standard error values are provided for the AUROC (or AUC). Moreover, the accuracy values tend to be much larger for the AD-Covid Gene Regulatory Network (GRN) model genes than for the benchmark (87.5% for Hippocampus, 81.25% for DLPFC, 85% for LTL, 82.5% for combined regions, 77.5% for benchmark). Please see **Supplementary Table A.6** for more details.

Figure A.17 Decision Curve Analysis (DCA) for Covid-19 Severity Prediction for Covid-19 Positive Patients in the 20 Testing Samples: ICU (Class 1, Severe) vs. Non-ICU (Class 0, Not Severe):



Here, there are 7 different options for strategies and DCA compares the Net Benefit for these options across the probability thresholds. These thresholds are the point where a Covid-19 positive individual would be predicted as severe and sent to the Intensive Care Unit (ICU) of the hospital. For a given Probability Threshold, the optimal option is the one with the highest Net Benefit model; this ensures that individuals select optimal options based on personal preferences (measured by the Probability Thresholds). Each 0.1 unit increase of Net Benefit refers to Net Benefit on the Treated (Covid-19 patients sent to the ICU) and represents that 1 out of 10 Covid-19 patients who are truly in a severe state are correctly predicted as being severe and sent to the ICU, holding the number of non-severe patients who are needlessly sent to the ICU constant. Thus, the Net Benefit helps us measure the advantage of a given model in terms of correctly sending severe patients to the ICU. The DLPFC model tends to have the highest Net Benefit for a majority of probability thresholds.

Figure A.18 Model Evaluation for the AD-Covid Logistic Regression (LR) Model and the AMP-AD LR Model for Predicting Alzheimer's Disease (AD) on 24 Human Samples in Testing Data:

Please note that we had 24 total cell-type human samples from the Superior Frontal Gyrus (SFG) brain region of the Frontal Cortex: 12 with AD and 12 Controls.

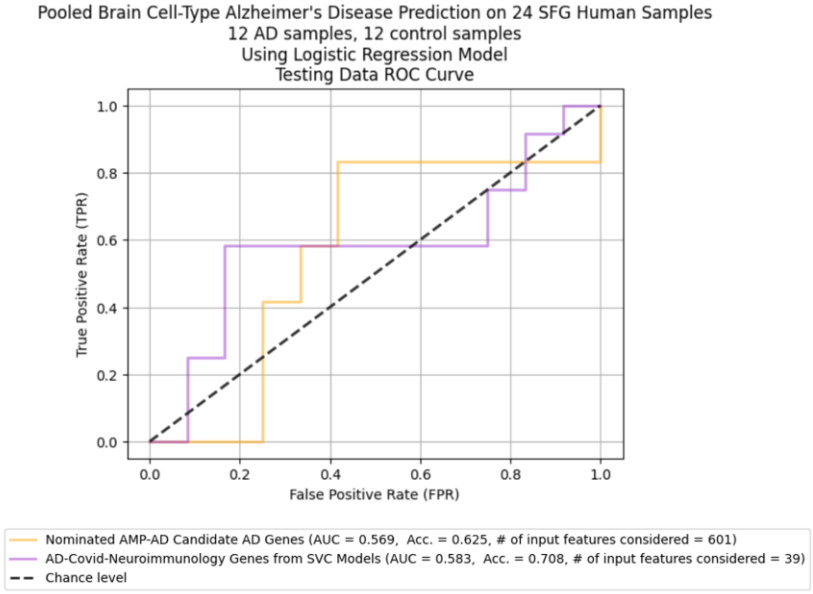


Figure A.18A) - Receiver Operating Characteristic (ROC) curve corresponding to both models and their respective area under curve (AUC) values for classifying AD (or control) in the 24 samples in testing data.

Based on these results, our 36 optimal AD-Covid genes (35 were found in the SFG gene expression data) used for predicting Covid-19 severity in the SVM models) are more predictive of AD than the AMP-AD genes are. We found the AMP-AD LR model has an AUC of 0.569 and accuracy of 62.5%, while our AD-Covid LR model has an AUC of 0.583 (slightly higher) and a much better accuracy of 70.8% on test data.

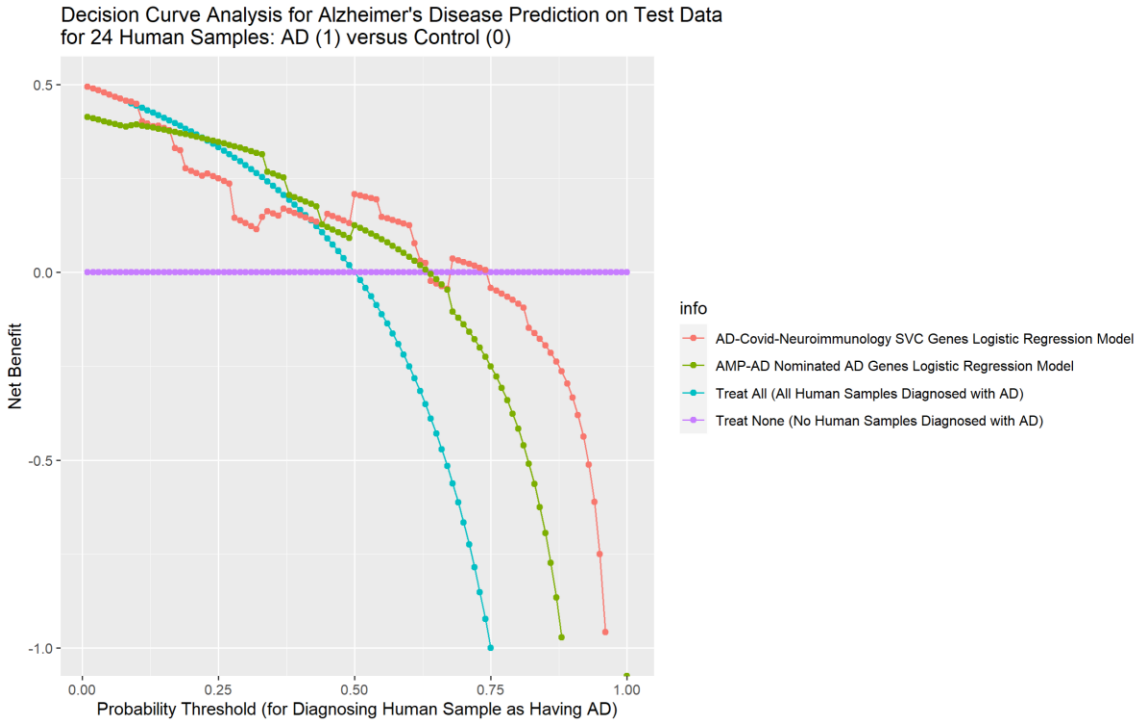
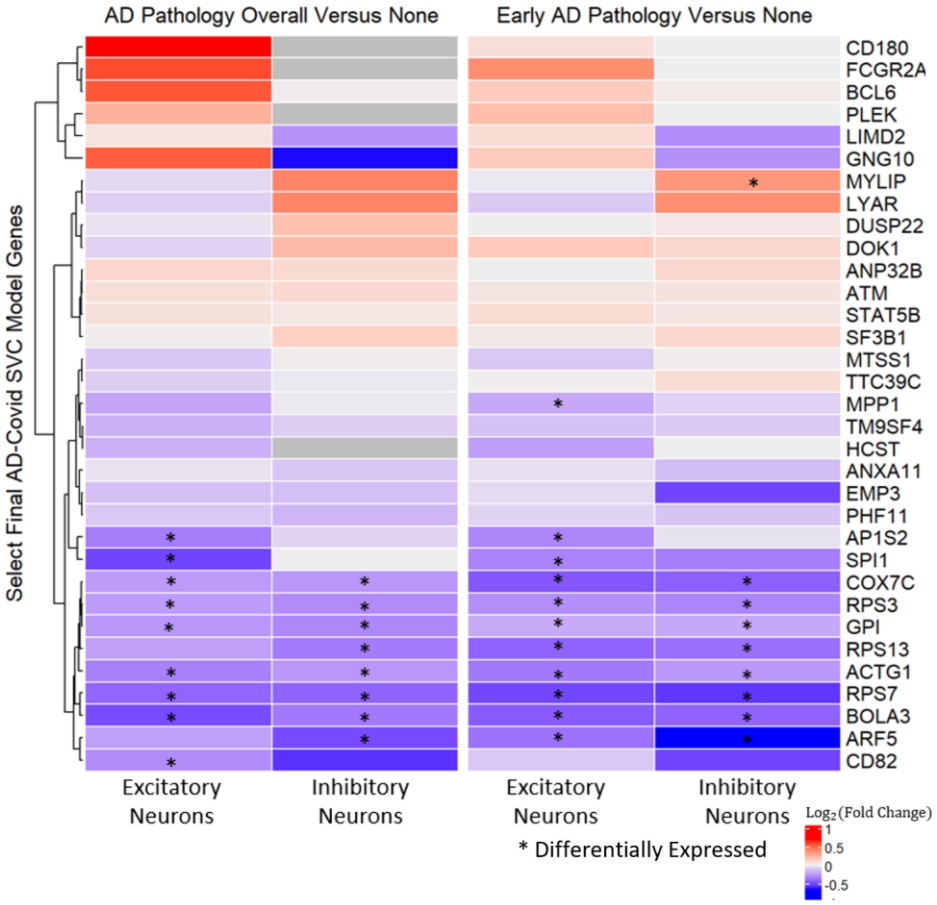


Figure A.18B) – There are 4 different options for strategies and Decision Curve Analysis (DCA) compares the Net Benefit for these options across the probability thresholds.

These thresholds are the point where a patient would be predicted (diagnosed) as having Alzheimer's disease (AD) and may presumably undergo life changes and/or treatments based on this diagnosis. Please note that based on the DCA theory, looking at the Net Benefit on the Treated (net benefit on individuals predicted as having AD, whether or not they actually have it) to select the optimal predictive model is analogous to looking at the Net Benefit on the Untreated (net benefit on individuals predicted as not having AD, whether or not they actually have it), or the Net Benefit Overall (Net Benefit on the Treated + Net Benefit on the Untreated). Thus, we could have looked at Net Benefit on the Untreated or Net Benefit Overall, but selected the default of Net Benefit on the Treated for simplicity (and consistency in interpretation compared with the Covid-19 severity DCA analysis). For a given Probability Threshold, the optimal option is the one with the highest Net Benefit model; this ensures that individuals select optimal options based on personal preferences (measured by the Probability Thresholds). Each 0.1 unit increase of Net Benefit refers to Net Benefit on the Treated (individuals predicted as having AD) and represents that 1 out of 10 individuals who truly have AD are correctly diagnosed as having AD (and receive the appropriate treatment as an AD patient), holding the number of control (non-AD) patients who are wrongly diagnosed with AD constant. Thus, the Net Benefit helps us measure the advantage of a given model in terms of correctly predicting an AD patient has AD. We see that the AD-Covid LR model tends to have the highest Net Benefit for more probability thresholds than the AMP-AD model does. **File A.7** contains more information.

Figure A.19 Differential Expression Analysis for AD-Covid genes from 4 AD-Covid GRN-based models (for predicting Covid-19 severity) on external single-cell transcriptomic data for Excitatory and/or Inhibitory Neurons for AD human samples versus Controls.



Heatmap using Excitatory Neuron and Inhibitory Neuron cell-type snRNA-seq data from Mathys et. al to analyze log2(fold change) of gene expression for 33 identified (out of 36) AD-Covid final SVC model genes (across 4 models) between 2 conditions: 1. AD pathology overall versus none (healthy controls), 2. Early AD pathology versus none. The heatmap goes from negative log fold change (blue: implying decreased gene expression during AD pathology) to positive (red: increased gene expression during AD) and (*) denotes the gene is Differentially Expressed (based on Individual Model).

Figure A.20 AD-Covid genes and regulatory networks for predicting Covid-19 severity in the Hippocampus and Lateral Temporal Lobe (LTL):

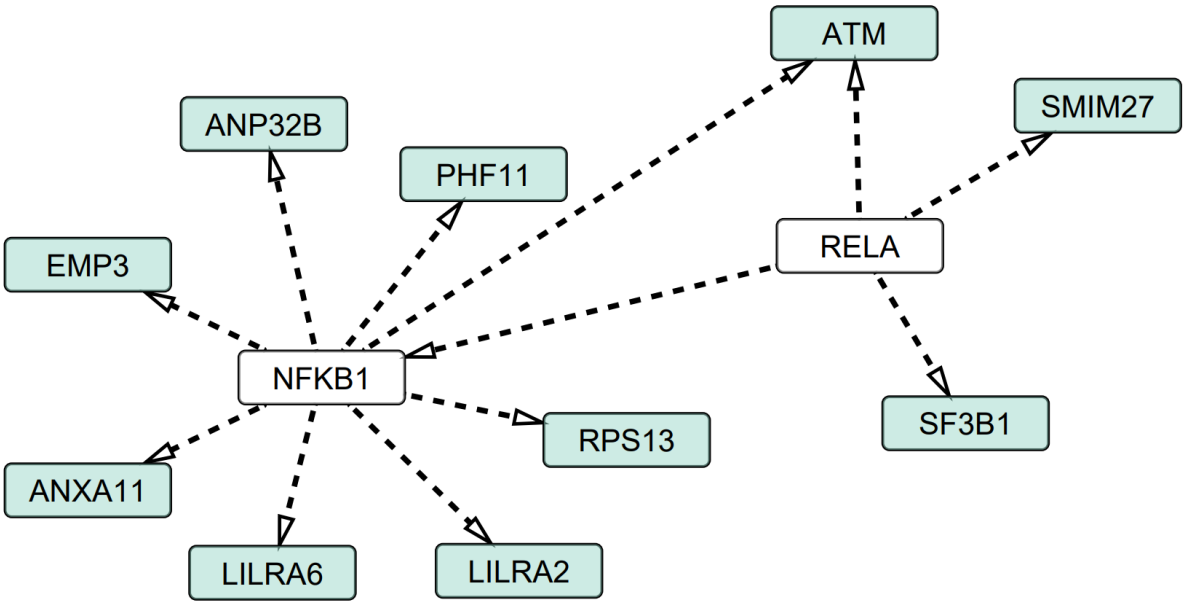


Figure A.20A) - Subnetwork of the Hippocampus gene regulatory network (GRN) relating to all 10 of the Hippocampus optimal genes for predicting Covid-19 severity (*N=10*) with AD-Covid shared genes.
Turquoise: genes/TFs found in the Hippocampus final model. White: AD-Covid shared genes. There was no overlap between both sets.

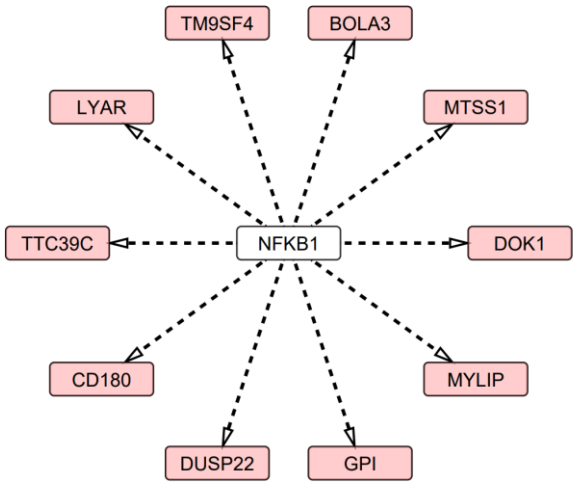


Figure A.20B) - Subnetwork of the LTL GRN relating to all of the 10 LTL optimal genes for predicting Covid-19 severity (*N=10*) with AD-Covid shared genes.
Pink: genes/TFs found in the LTL final model. White: AD-Covid shared genes. There was no overlap between both sets.

Figure A.21 Additional SNP Regulatory Network Examples

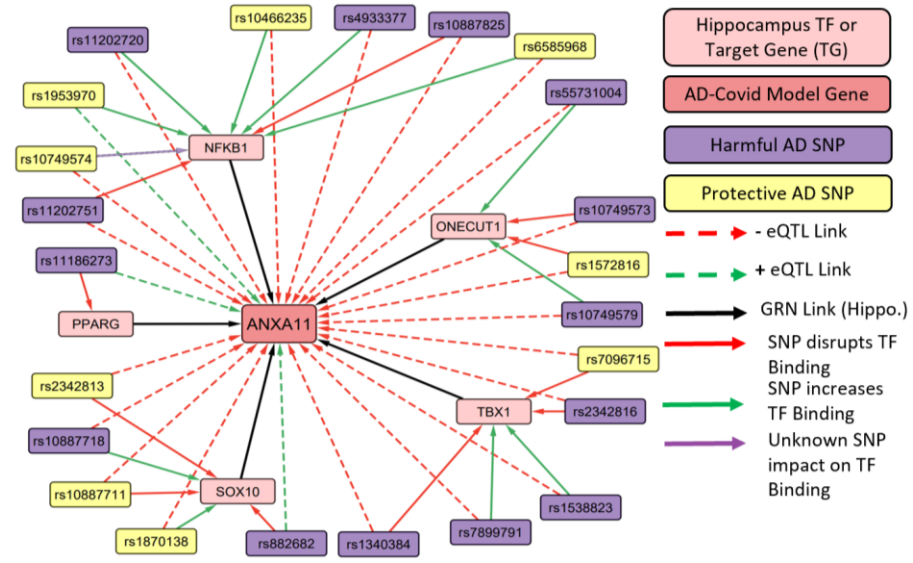


Figure A.21A) - This subnetwork links AD GWAS SNPs to the Hippocampus GRN, focusing on the regulation of ANXA11, a final AD-Covid gene in the Hippocampus.

Regulation of *ANXA11* by 5 TFs is impacted by various AD SNPs (located directly within regulatory elements), which also has eQTL validation from various brain cell-types (linking the SNPs causally with changes in *ANXA11* expression). Red dashed arrows show negative expression quantitative trait loci (eQTL) slope between SNP and *ANXA11*, while green dashed arrows show a positive eQTL slope (SNP associated with increased expression of target gene *ANXA11*).

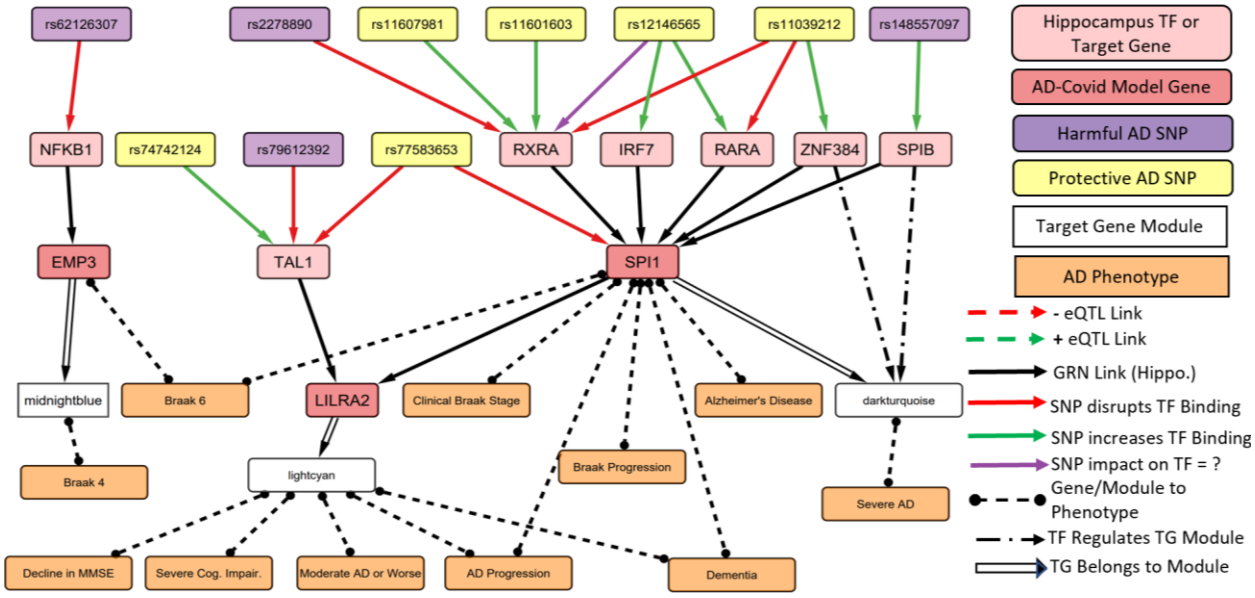


Figure A.21B) – Here, we highlight AD SNPs that may dysregulate Hippocampal expression of 3 of the optimal AD-Covid genes*: EMP3, LILRA2, SPI1*.

The legend on the right corresponds to the nodes and edges in this figure in the Hippocampus CA1 region. Protective (yellow) SNPs are associated with Controls and are associated with decreased AD risk while

harmful (purple) SNPs are associated with AD patients and associated with increased AD risk. SNPs that increase TF Binding (possibly by creating new motif) are green solid arrows while SNPs that disrupt TF Binding are red solid arrows; unknown SNP effects are purple solid arrows. Dark red corresponds to 1 of the 46 final AD-Covid SVC Model Genes. This subnetwork in the Hippocampus connects AD SNPs (found directly within regulatory elements) with TFs and TGs along with the modules the TGs belong to and associated phenotypes for the TGs and the modules. Some TFs may significantly regulate gene modules. Harmful AD rs62126307 disrupts NFKB1 binding to EMP3 (up-regulated in glioma tissues, associated with Braak 6 and a Braak 4 stage module) enhancer. AD SNPs disrupt regulation of SPI1 by 5 TFs (ZNF384 and SPIB TFs regulate SPI1's Severe AD module). SPI1 regulates immune functions in AD and genes upregulated in microglia, is strongly correlated with AD, Braak Progression, and involved in microglia-mediated AD neurodegeneration SNP rs77583653 disrupts the ability of TAL1 and SPI1 to regulate LILRA2 (associated with worse AD).

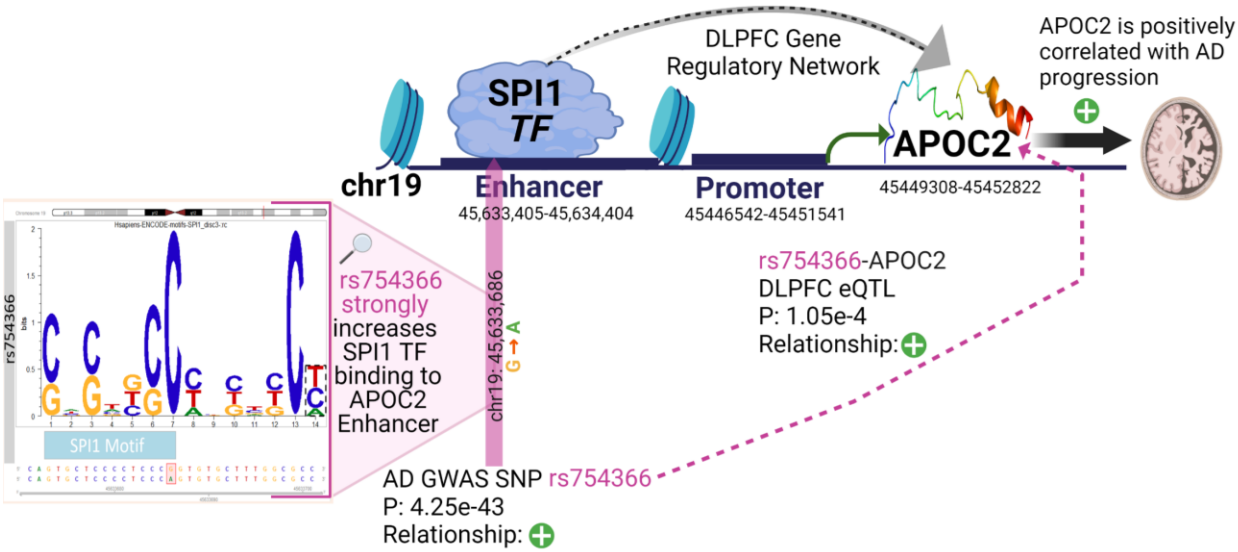


Figure A.21C) –Highly significantly harmful AD SNP rs754366 may increase TF binding of activator TF SPI1 to the *APOC2* enhancer in the DLPFC.

This figure zooms in to show how the SNP impacts the motif binding site. Increased APOC2 expression is associated with AD, APOE4, and its cognitive impairment module is enriched for Covid-19, neuron death, neuroinflammation, abnormal innate immunity, TNFa signaling via NFKB, brain death. There is eQTL validation in the Prefrontal Cortex to support this finding.

Figure A.22 SNP rs56344893 Potentially Found to Disrupt Regulation of KCNN4 in Hippo. & LTL

KCNN4 (also known as KCa3.1) is a Potassium Calcium Activated Channel that is found in cells such as the Microglia, the resident macrophage immune cell of the brain. Calcium binding to KCNN4 leads to the efflux of Potassium from the cell, which is key since Calcium signaling is disrupted during AD.

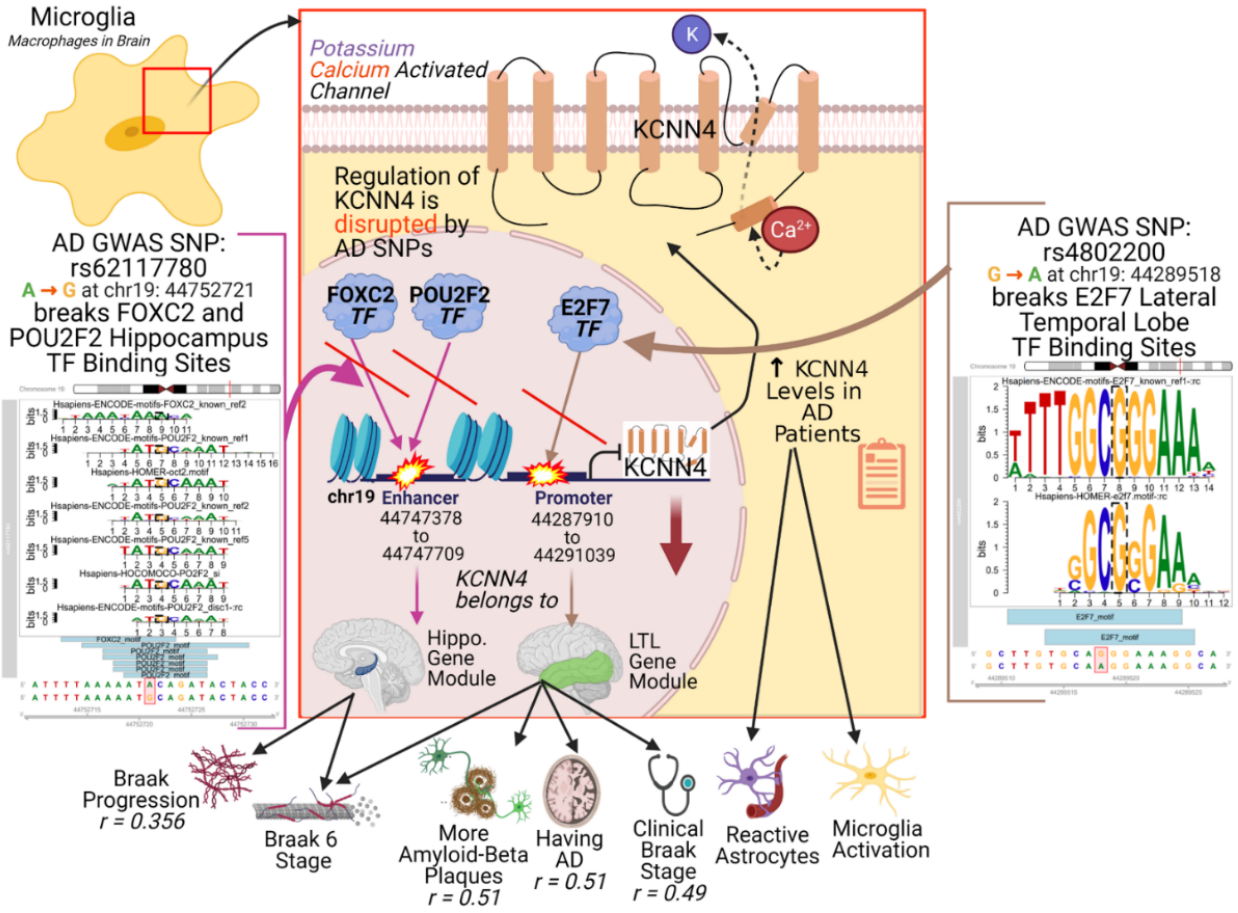
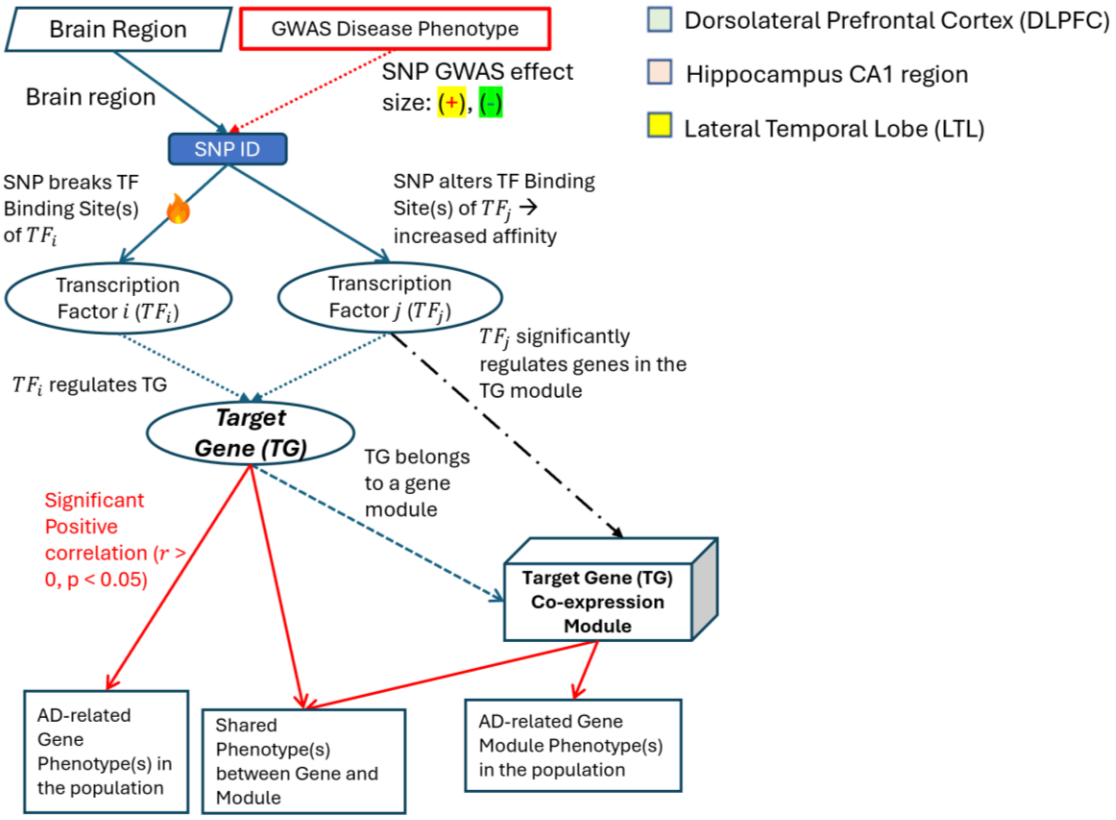


Figure A.22A) – predicted impact of various SNPs on the binding of TFs to KCNN4 regulatory regions in the Hippocampus and Lateral Temporal Lobe (LTL).

Here, SNP rs62117780 in the Hippocampus (changes the DNA base from an A to a G on chromosome 19 at hg19 position 44752721) disrupts the binding of FOXC2 and POU2F2 TFs to KCNN4's enhancer in the Hippocampus. Another SNP rs4802200 in the LTL (changes the DNA instead from a G to an A at hg19 position 44289518) and disrupts the binding of E2F7 to KCNN4's promoter in the LTL. Both SNPs also disrupt the binding of other TFs that regulate other genes (not shown). Thus, since the respective TFs are unable to bind to these regulatory regions for KCNN4 in the Hippocampus and LTL, they are unable to properly regulate KCNN4. We think these TFs may serve as repressors of KCNN4; hence, these non-coding SNPs (along with many we showed in provided in our SNP-effected-GRN and webtool) may explain the elevated levels of KCNN4 in AD patients (and why KCNN4 remains a strong drug target for AD), despite the lack of known SNPs in the KCNN4 gene-coding region. (Nonetheless, more follow-up analysis would be needed). These elevated levels are typically associated with reactive astrocytes and microglia activation and worse AD neurotoxicity, neuroinflammation, and neuropathology. We also found KCNN4 belongs to Hippocampus and LTL gene modules that are associated with worse AD outcomes. For instance, the Hippocampus gene module is positively associated with Braak Progression (r = 0.356).

Figure A.22B) – We zoom in on the epigenetic landscape of Chromatin Binding for AD SNPs Related to KCNN4 Regulation in microglia. We looked at the brain cell types of Microglia. We used the WashU EpiGenome Browser (Li et. al 2022) webtool with the Corces_scATAC_BroadCellTypes (Corces et al. 2020) tracks. This figure zooms in on SNP rs62117780.

Figure A.23 Interpreting SNP-effected-GRNs: Regulatory Network Linking Variants to Phenotypes.



GWAS SNPs altering the ability of Transcription Factors (TFs) to bind to TF Binding Sites (TFBSs) to regulate their target genes. This is an example of our predicted SNP-effected-GRNs. This figure explains how to interpret our Regulatory Network Linking Variants to Phenotypes. Boxes shaded in light green represent the DLPFC region, those in light pink represents Hippocampus, and those in light yellow represent LTL. Here, we take the SNP IDs and see how they alter TF Binding Sites (TFBS) in any of these 3 Brain Regions, which then impact the binding of TFs and regulation of other genes. Otherwise, if the TF-TG relationship is unknown (not in TRRUST2), it is not mentioned in the network. Each TG belongs to a TG co-expression module, and that Target Gene Module may be significantly positively ($r > 0$, $p < 0.05$) associated with various AD phenotypes in particular brain regions. Moreover, genes can be significantly positively associated with their own AD phenotypes. The gene and module may also share some common phenotypes together. All in all, we see that the SNP in the TF Binding Sites can strongly or weakly disrupt the TF to Target Gene relationship such that this link may be broken. If a TF significantly regulates the TG module (based on RTN findings), please note that a black arrow will be drawn from the TF to the Target Gene module.

Figure A.24 Some SNP-effected-GRN visualizations based on **Figure A.23.**

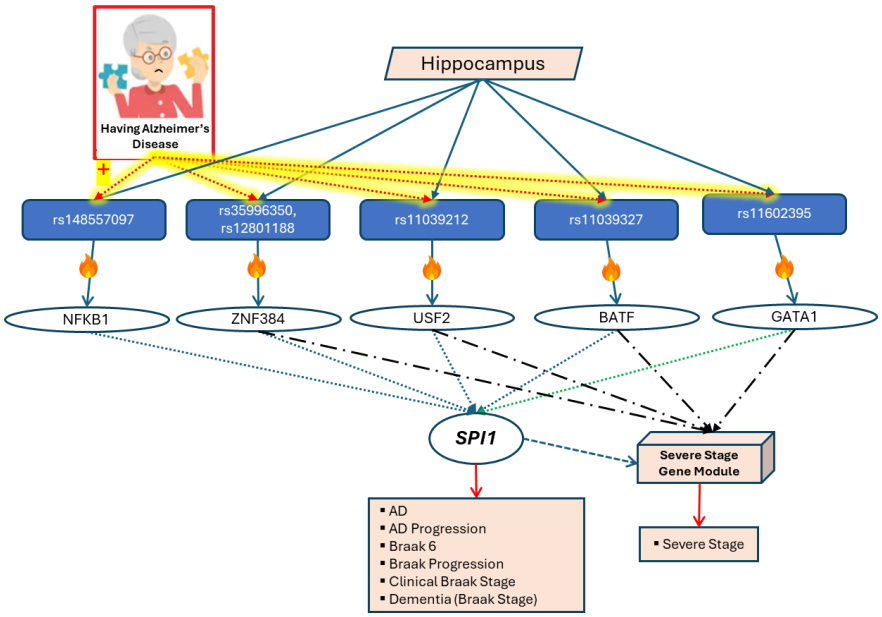


Figure A.24A) Some SNPs that Disrupt Regulation of *SPI1* in the Hippocampus.

There are several SNPs that disrupt *SPI1* regulation in the Hippocampus. Here, we focus on 5 SNPs that disrupt the regulation of *SPI1* in the Hippocampus by 2 TFs: RXRA and RARA. We also include another TF, NFKB1, that regulates *SPI1* and 2 SNPs that impact NFKB1's ability to properly regulate *SPI1*. We found that *SPI1* is a key Transcription Factor (TF) that is associated with regulation of microglia and may be a key driver in AD onset and progression. In the Hippocampus, *SPI1* belongs to a Severe Stage module, where it is positively associated with the Severe Stage. It is also positively correlated with AD, AD Progression, Braak 6, Braak Progression, Clinical Braak Stage, and Dementia (based on the Braak Stage).

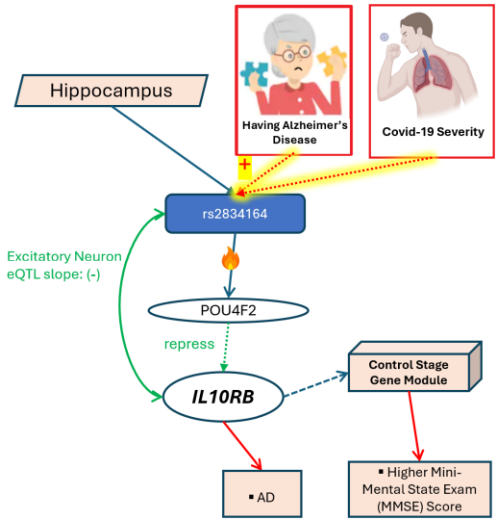


Figure A.24B) Some SNPs that alter Regulation of *IL10RB* in the Hippocampus:

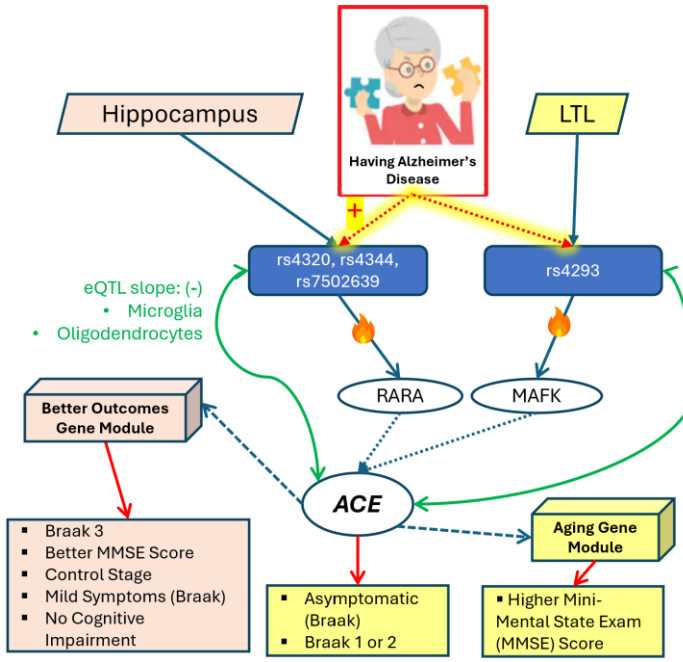


Figure A.24C) Some SNPs that alter regulation of target gene *ACE* in the Hippocampus and LTL.

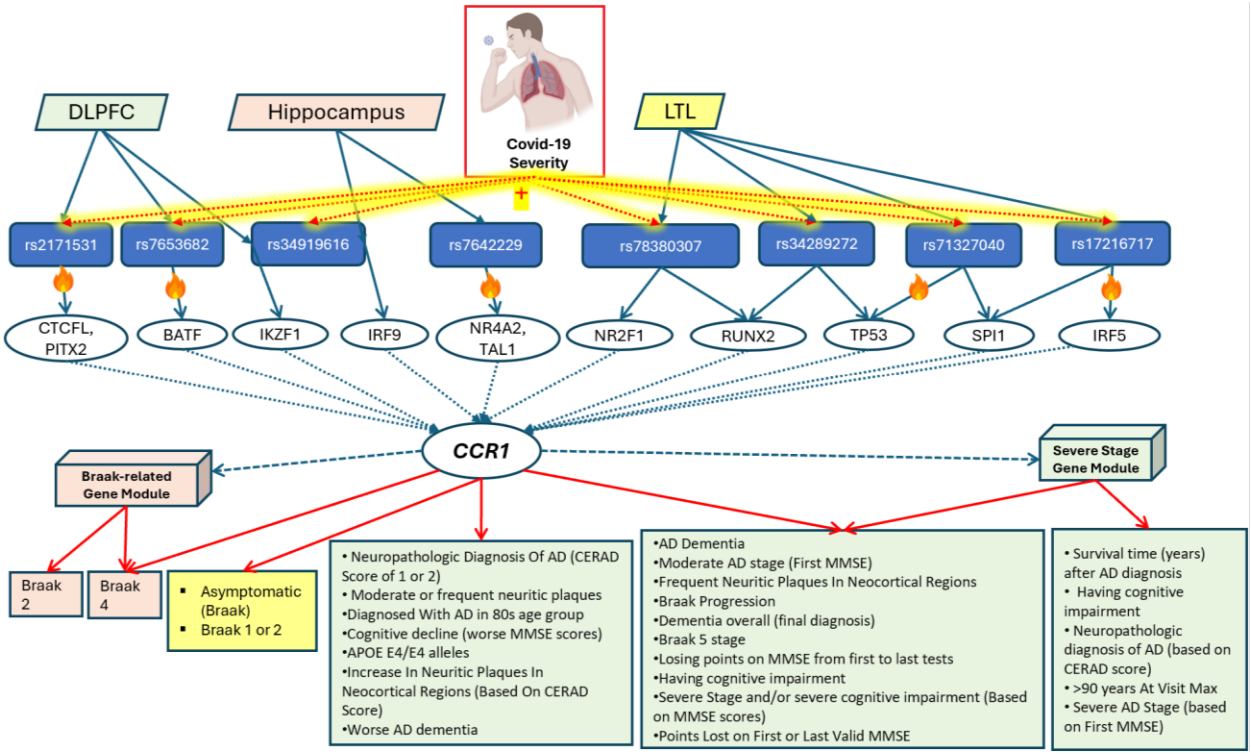


Figure A.24D) Some SNPs that alter Regulation of CCR1 in the Hippocampus, DLPFC, and LTL.

Here, there are various Covid-19 severity risk SNPs that impact regulation of target gene *CCR1* in all the 3 brain regions. *CCR1* is associated with different population phenotypes across the brain regions.

Figure A.25 SNP rs3851178: alters TF Binding to TF Binding Sites in All 3 Brain Regions (Hippocampus, LTL, and DLPFC)

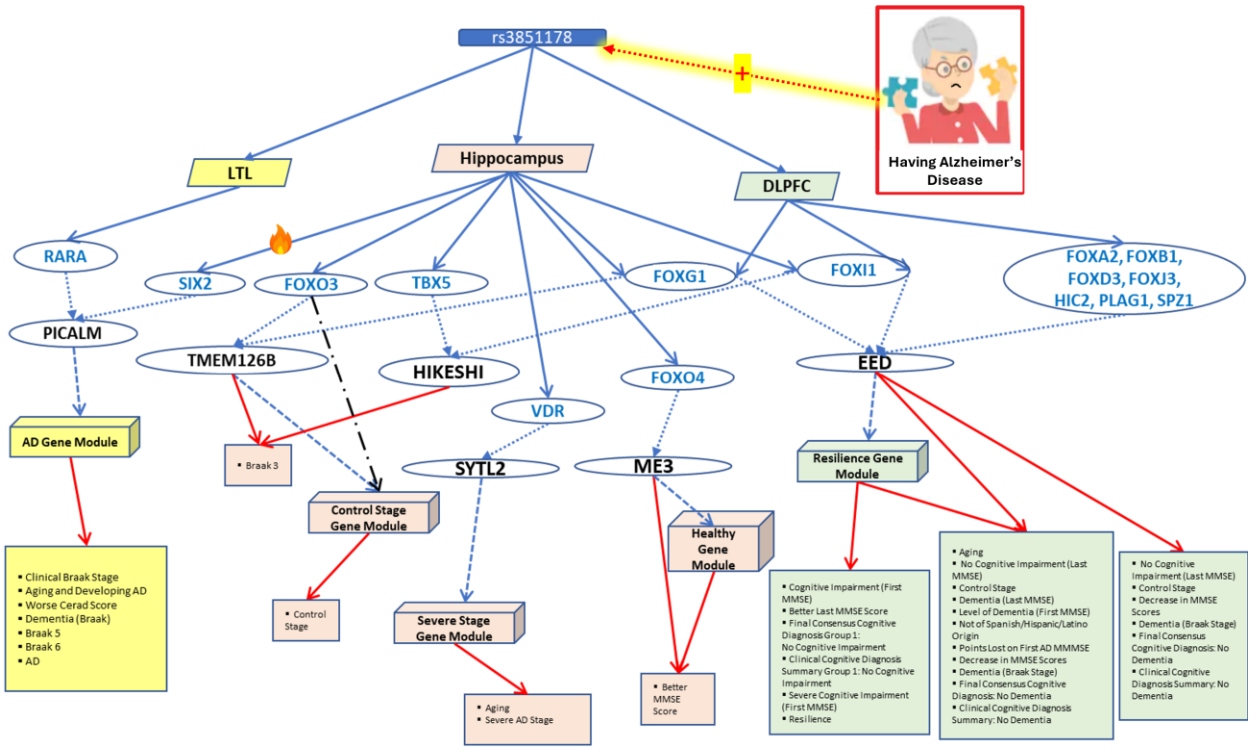
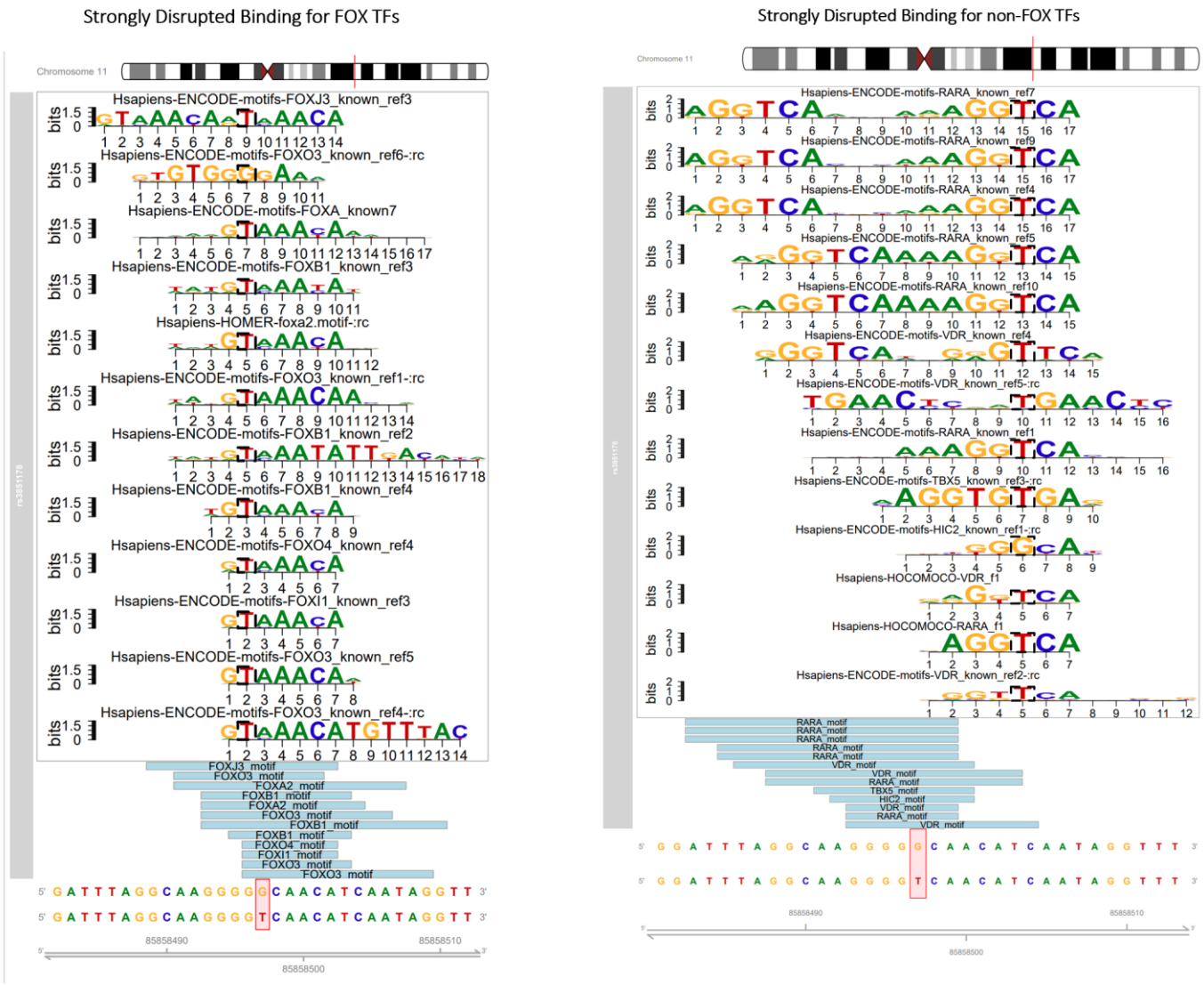


Figure A.25A) Impact of rs3851178 on all 3 brain regions is shown in this SNP Regulatory Pathway. Here, this SNP impacts RARA's regulation of *PICALM* in the LTL and SIX2's regulation of PICALM in the Hippocampus. *PICALM* belongs to an AD gene module in the LTL. This SNP also impacts the regulation of TMEM126B in the Hippocampus, and this gene is associated with the Control Stage Module and with Braak 3. FOXO3, which regulates TMEM126B, also regulates the Control Stage Module. In addition, this SNP impacts the regulation of Severe Stage module gene *SYTL2* and Healthy Gene Module gene *ME3* in the Hippocampus. Furthermore, this SNP impacts regulation of Resilience gene module gene *EED* by many Transcription Factors (TFs); *EED* and its associated module are positively correlated with improvements in cognitive impairment outcomes for individuals who have AD. We found that FOXG1 and FOXI1 are 2 TFs whose regulation in the Hippocampus and DLPFC are impacted by this SNP.



SNP rs3851178 (p < 5.5e-15)

Figure A.25B) - Motifs Broken by a SNP Found in All 3 Brain Regions: rs3851178.
Please note that this Figure focuses on rs3851178, whose regulatory impacts are illustrated in the previous **Figure A.24**. SNP rs3851178 is a change from a G to a T base in position 85858497 (hg19) of chromosome 11, which impacts the regulation of these 15 TFs: FOXA2, FOXB1, FOXD3, FOXG1, FOXI1, FOXO3, FOXO4, HIC2, PLAG1, RARA, SIX2, SPZ1, TBX5, VDR. Disrupted genes (not shown), are: PICALM (by RARA and SIX2), TMEM126B (by FOXO3 and FOXG1), HIKESHI (by TBX5 and FOXI1), EED (by FOXG1 and FOXI1), SYTL2 (by VDR), ME3 (by FOXO4).

Figure A.26 Example of SNP for regulating PPP1R37

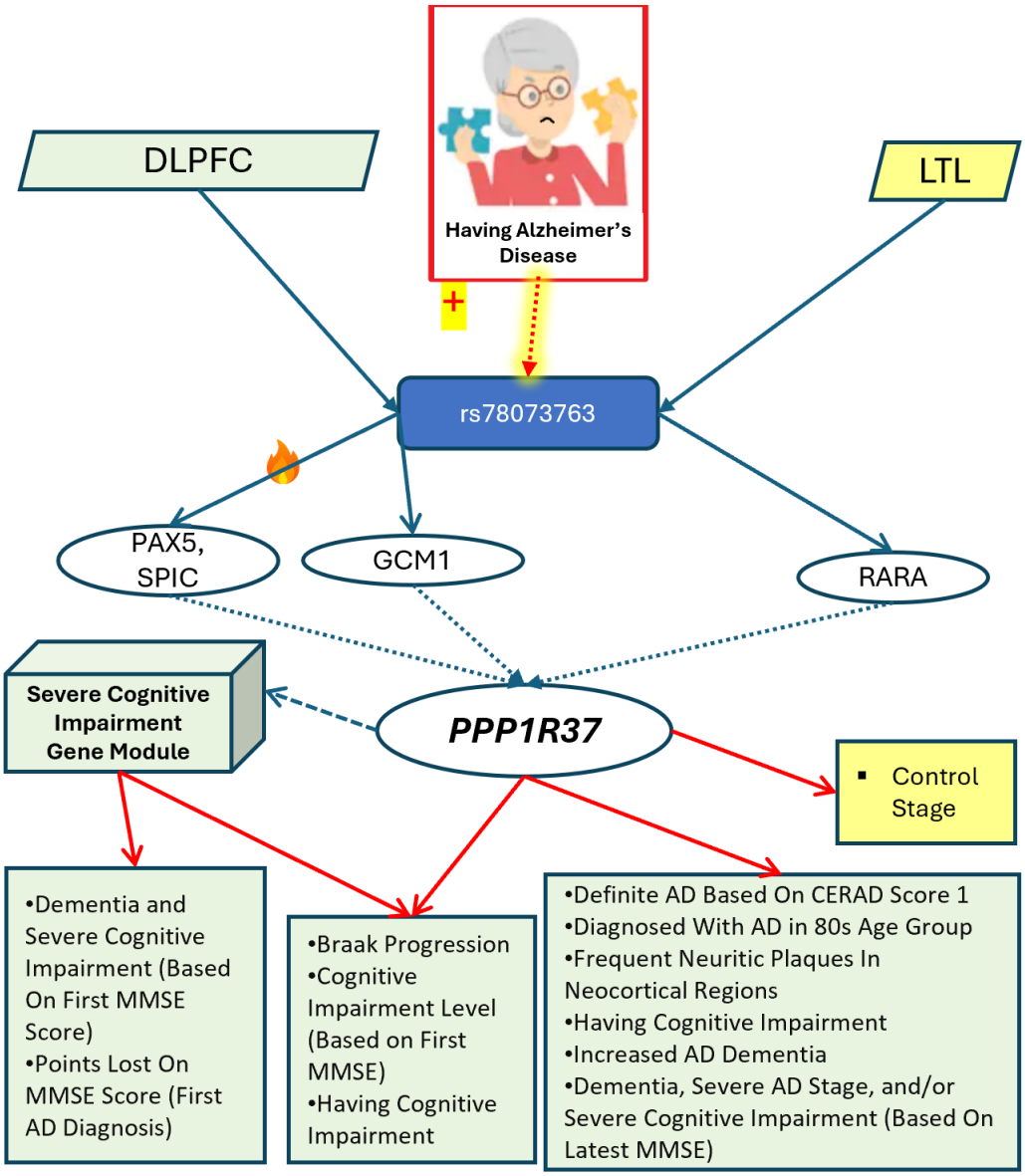


Figure A.26A) – SNP altering regulation of TG *PPP1R37* in the LTL and DLPFC. Alzheimer's disease (AD) risk SNP rs78073763 impacts regulation of *PPP1R37* in both brain regions. In the DLPFC, this SNP disrupts binding of PAX5 and SPIC TFs to regulatory elements (REs) for this TG but boosts binding of GCM1 instead. RARA binding to an RE for *PPP1R37* increases in the LTL. This gene is assigned to a gene co-expression module in the DLPFC that is associated with severe cognitive impairment. Moreover, this TG in the DLPFC is associated with worsening AD-related phenotypes, while it has higher expression in the LTL for Control humans instead.

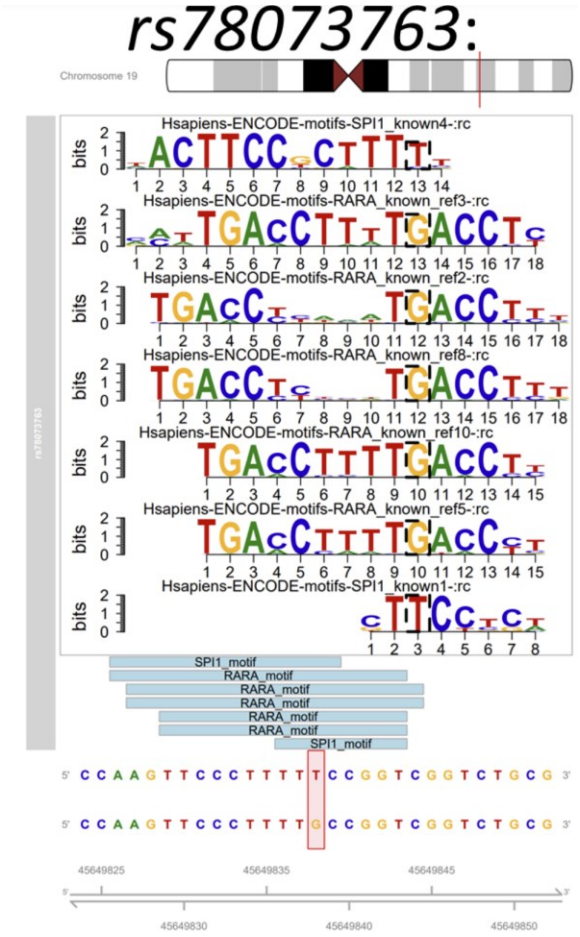


Figure A.26B) – SNP rs78073763 alters multiple possible binding sites of RARA in the LTL.

Figure A.26C) – Epigenetic landscape of Chromatin Binding for AD SNP rs78073763 (chr19:45649838, hg19 position).

We looked at Microglia, the cell type associated with the resident macrophage immune cells in the brain. We used the WashU EpiGenome Browser (Li et. al 2022) with the Corces_scATAC_BroadCellTypes (Corces et al. 2020) tracks. We zoomed in on rs78073763.

§ A.3 Supplementary tables

Table A.1: Resources used for Gene enrichment and annotation

Sources:	**Enrichment Information**
Metascape (metascape.org) (Zhou et al. 2019a)	● KEGG Pathway Analysis ● Gene Ontology Biological Processes ● Reactome Gene Sets ● Canonical Pathways ● CORUM (Comprehensive Resource of Mammalian Protein Complexes) ● TRRUST ● Transcription_Factor_Targets ● COVID ● DisGeNET (Disease Gene Network) ● PaGenBase (Pattern Gene Database) ● Significantly-enriched mCode (Minimal Common Oncology Data Elements) Protein-Protein Interactions (PPIs)
Gprofiler (biit.cs.ut.ee/gprofiler/gost)(Reimand et al. 2007)	● Gene Ontologies ● KEGG Pathways ● Transcription Factors ● Reac ● MIRNA ● HPA ● CORUM (Comprehensive Resource of Mammalian Protein Complexes) ● HP (Human Phenotypes) ● WikiPathways analysis
BaderLab *Human Entrez Gene* (Bader et. al, 2021)	● Gene Ontologies (Molecular Function, Cellular Components, Biological Processes) ● All Pathways ● Drug Targets (DrugBank: small molecules, nutraceutical, illicit, experimental, and/or approved) ● Disease Phenotypes ● Transcription Factors (TFs; Human) ● miRs ● Molecular Signatures Database ● Institute of Bioinformatics (IOB) Pathways ● Reactome Pathways ● NetPath Pathways ● Panther Pathways ● NCI Nature Pathways ● WikiPathways ● HumanCyc Pathways ● KEGG Pathways
Maayan Lab (https://maayanlab.cloud/Harmonizo	Please note that these 125 Different Data Sources were used from Maayan Lab: ● Achilles Cell Line Gene Essentiality Profiles ● Allen Brain Atlas Adult Human Brain Tissue Gene Expression Profiles ● Allen Brain Atlas Adult Mouse Brain Tissue Gene Expression Profiles

me/download)(Ro uillard et al. 2016)	● Allen Brain Atlas Developing Human Brain Tissue Gene Expression Profiles by Microarray and/or RNA-seq ● Allen Brain Atlas Prenatal Human Brain Tissue Gene Expression Profiles ● BIND Biomolecular Interactions ● Biocarta Pathways ● BioGPS Cell Line Gene Expression Profiles ● BioGPS Human and/or Mouse Cell Type and Tissue Gene Expression Profiles ● BioGRID Protein-Protein Interactions ● CCLE Cell Line Gene CNV Profiles, Gene Expression Profiles, and/or Gene Mutation Profiles ● CHEA Transcription Factor Binding Site Profiles and/or TF Targets ● ClinVarSNP-Phenotype Associations ● CMAP Signatures of Differentially Expressed Genes for Small Molecules ● Combined Pathways Pathways ● COMPARTMENTS Curated, Experimental, and/or Text-mining Protein Localization Evidence Scores ● CORUM Protein Complexes ● COSMIC Cell Line Gene CNV and/or Cell Line Gene Mutation Profiles ● CTD Gene-Chemical Interactions and/or Gene-Disease Associations ● dbGAP Gene-TraitAssociations ● DEPOD Substrates of Phosphatases ● DIP Protein-Protein Interactions ● DISEASES Curated, Experimental, and/or Text-mining Gene-Disease Association Evidence Scores ● DrugBank Drug Targets ● ENCODEHistoneModification Site Profiles ● ENCODE Transcription Factor Binding Site Profiles and/or TF Targets ● ESCAPEOmics Signatures of Genes and Proteins for Stem Cells ● GAD Gene-Disease Associations and/or High Level Gene-Disease Associations ● GDSC Cell Line Gene Expression Profiles ● Gene RIF Biological Term Annotations ● GenesigDB Published Gene Signatures ● GEO Signatures of Differentially Expressed Genes for DISEASES, Gene Perturbations, Kinase Perturbations, Small Molecules, Viral Infections, and/or Transcription Factor Perturbations ● GO Annotations: Biological Process, Cellular Component, and/or Molecular Function ● GTExeQTL ● GTEx Tissue Gene Expression Profiles and/or Tissue Sample Gene Expression Profiles ● Guide to Pharmacology Chemical Ligands of Receptors ● Guide to Pharmacology Protein Ligands of Receptors ● GWAS Catalog SNP-Phenotype, dbSNP-Disease and/or dbSNP-Phenotype Associations ● Heiseretal., PNAS, 2011 Cell Line Gene Expression Profiles ● HMDB Metabolites of Enzymes ● HPA Gene Expression Profiles: Cell Line, Tissue, Tissue Sample ● HPA Tissue Protein Expression Profiles ● HPM Cell Type and Tissue Protein Expression Profiles

	● HPO Gene-Disease Associations ● HPRD Protein-Protein Interactions ● Hub Proteins Protein-Protein Interactions ● HuGE Navigator Gene-Phenotype Associations ● Human Cyc Biomolecular Interactions and/or Cyc Pathways ● IntAct Biomolecular Interactions ● InterPro Predicted Protein Domain Annotations ● JASPAR Predicted Transcription Factor Targets ● KEA Substrates of Kinases ● KEGG Biomolecular Interactions and/or Pathways ● KinativKinaseInhibitorBioactivity Profiles ● KinomeScan Kinase Inhibitor Targets ● Klijnetal. (Nat.Biotechnol., 2015 Cell Line Gene CNV Profiles, Gene Expression Profiles, Cell Line, Gene Mutation Profiles) ● LINCSL1000 CMAP Signatures of Differentially Expressed Genes for Small Molecules ● LOCATE Curated and/or Predicted Protein Localization Annotations ● MiRTarBasemicroRNA Targets ● MotifMap Predicted Transcription Factor Targets ● MPO Gene-Phenotype Associations ● MSigDB Cancer Gene Co-expression Modules ● MSigDB Signatures of Differentially Expressed Genes for Cancer Gene Perturbations ● NURSA Protein-Protein Interactions and/or Protein Complexes ● OMIM Gene-Disease Associations ● PANTHER: Biomolecular Interactions, Pathways ● Pathway Commons Protein-Protein Interactions ● Phospho Site Plus Phosphosite-Disease Associations and/or Substrates of Kinases ● Phosphosite Text-mining Biological Term Annotations ● PID Biomolecular Interactions and/or Pathways ● ProteomicsDB Cell Type and Tissue Protein Expression Profiles ● Reactome Biomolecular Interactions and/or Pathways ● ReconX Predicted Biomolecular Interactions ● Roadmap Epigenomics Cell and Tissue DNA Methylation and/or Gene Expression Profiles ● Roadmap Epigenomics Histone Modification Site Profiles ● SILAC Phosphoproteomics Signatures of Differentially Phosphorylated Proteins for Drugs, Gene Perturbations, and/or Protein Ligands ● Targets can Predicted Conserved and/or Non-conserved microRNA Targets ● TCGA Signatures of Differentially Expressed Genes for Tumors ● TISSUES Tissue Protein Expression Evidence Scores: Curated, Experimental, Text-mining ● TRANSFAC Curated and/or Predicted Transcription Factor Targets ● VirusMINT Protein-Viral Protein Interactions and/or Protein-Virus Interactions ● Wiki Pathways Pathways

WGCNA Package (Langfelder and Horvath 2008)	● Different brain lists from several studies, brain region markers (from human.brain-map.org (Allen Brain Atlas Data Portal)), CHDI-based lists on the neurodegenerative Huntington's Disease, Blood Atlas lists, Stem Cell lists, and Immune Pathway lists.
ABAEnrichment Package (Grote et al. 2016)	● Allen Brain Atlas enrichment information for different regions in the brain during 5 stages of development (prenatal, infant, child, adolescent, and adult)
Psygenet2r Package (Gutiérrez-Sacristán et al. 2017)	● Enrichment information on psychiatric diseases such as: bipolar disorder, schizophrenia, substance-induced depressive disorder, and psychoses from PsyGeNET (Psychiatric disorders Gene association NETwork).
TissueEnrich Bioconductor Package(Jain and Tuteja 2021)	● Enrichment information on tissue-specific genes
ClusterProfiler(Yu), DOSE(Yu et. al 2014), and msigdbr(Bhuva et. al 2021) Bioconductor Packages	● Gene Ontology Analysis ● Disease enrichment analysis using the Disease Ontology (DO), Network of Cancer Genes (NCG), and DisGeNet (DGN). ● Medical Subject Headings (MeSH) Module-Level Enrichments using data from the gendoo and gene2pubmed sources for categories: Anatomy (A); Diseases (C); Analytical, Diagnostic, and Therapeutic Techniques and Equipment (E); Psychiatry & Psychology (F); Biological Sciences (G). ● Reactome Pathway Analysis ● WikiPathways analysis ● Cell Marker Analysis ● Molecular Signatures Database (MSigDB) analysis for all 8 major collections (H, C1 to C7). ● KEGG pathway and KEGG module analysis.
rentrez package(Lê Cao et al. 2011; Hastie 2020)	● NCBI gene summary information
Pathview Bioconductor Package(Luo and Brouwer 2013b)	● Visualize data-driven molecular Kegg pathway graphs.

Table A.2: Breakdown of Human Cell-Type Samples for the Superior Frontal Gyrus (SFG) used for Logistic Regression Models for Predicting Alzheimer's Disease (AD) or Not (Control)

Breakdown of Human Cell-Type Samples for Superior Frontal Gyrus (SFG)						
	Endothelial	**Myeloid (Microglia)**	**Astrocyte**	**Neuron**	**TOTAL**	**Group**
Total	27	20	19	42	108	*Overall*
Control	17	12	12	21	62	
AD	10	8	7	21	46	
Total	6	6	6	6	24	*Testing Data*
Control	3	3	3	3	12	
AD	3	3	3	3	12	
Total	21	14	13	36	84	*Training Data*
Control	14	9	9	18	50	
AD	7	5	4	18	34	

Table A.3: Breakdown of SNPs with by P-value from Major Alzheimer's Disease (AD) and Covid-19 Severity Genome-Wide Association Studies (GWAS):

Breakdown of SNPs by P-Value Across AD and Covid-19 Severity GWAS				
Study Name	More information	# of SNPs with P-value < 5e-5	# of SNPs with P-value < 5e-8	Citation
Kunkle Phase 1 2019		5	0	Kunkle, B. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet* **51**, 414–430 (2019).
Jansen 2019		7,631	2,391	Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet* **51**, 404–413 (2019).
Wightman 2021		11,200	2,709	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.

Bellenguez 2022		17,781	5,635	New insights into the genetic etiology of Alzheimer's disease and related dementias | Nature Genetics. https://www.nature.com/articles/s41588-022-01024-z.
PanUK Biobank	Phenotypes: Dementia in AD (icd10-F00) and AD (icd10-G30 and phecode290).	2,307 (icd10-F00: 762 SNPs, Phecode290: 764 SNPs, icd10-G30: 781 SNPs)	0	Pan UKBB | Pan UKBB. https://pan.ukbb.broadinstitute.org/
Covid-19 Hg-Round 7 Meta-analysis 2022	Covid-19 Positive all populations: hospitalized versus non-hospitalized	1,642	397	https://www.covid19hg.org/results/r7/
Total		28,597	7,117	

The first 4 studies are used for Alzheimer's disease (AD) Single Nucleotide Polymorphisms (SNPs) and the last study (last row) is for Covid-19 severity SNPs.

Table A.4: Data Resources for eQTLs (Linking SNPs to Changes in Target Gene Expression)

Please note that when the raw (original) eQTL data was provided, we instead filtered for p-value < 5e-3. Otherwise, we used the original filtered eQTL p-values from the study (that may be stricter with a more stringent p-value cutoff).

Data Resources for Expression Quantitative Trait Loci (eQTL): SNP to Target Gene Links		
Region	**More Details**	**Source**
Brain Cell Types	Excitatory Neurons, Inhibitory Neurons, Microglia, Oligodendrocytes	Zeng, B. *et al.* Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits. *Nat Genet* **54**, 161–169 (2022).
Brain Cell Types	Excitatory Neurons, Inhibitory Neurons, Microglia, Oligodendrocytes, Oligodendrocyte Precursor Cells (OPCs), Endothelial Cells, Pericytes, Astrocytes	Bryois, J. *et al.* Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat Neurosci* **25**, 1104–1112 (2022).
GTEx Version 8 Brain Tissues Significant eGenes	Caudate Basal Ganglia, Nucleus Accumbens, Cortex, Putamen Basal Ganglia, Frontal Cortex BA9, Cerebellum, Hippocampus, Hypothalamus, Anterior Cingulate	Genotype-Tissue Expression (***GTEx***) project. Single-Tissue cis-QTL Data: GTEX_Analysis_v8_eQTL.tar

	Cortex BA9, Spinal Cord Cervical C-, Cerebellar Hemisphere, Substantia Nigra	
Prefrontal Cortex	Brain Region	Resource.PsychEncode. http://resource.psychencode.org/. : Full_hg19_cis-eQTL
ROSMAP Brain Tissue	Brain eQTLs and cell-type eQTLs	Patel, D. *et al.* Cell-type-specific expression quantitative trait loci associated with Alzheimer disease in blood and brain tissue. *Transl Psychiatry* **11**, 1–17 (2021).
Framingham Heart Study (FHS) Blood Tissue	Blood eQTLs and cell-type eQTLs	Patel, D. *et al.* Cell-type-specific expression quantitative trait loci associated with Alzheimer disease in blood and brain tissue. *Transl Psychiatry* **11**, 1–17 (2021).

Table A.5: Full Alzheimer's Disease (AD) Gene Regulatory Network (GRN): Transcription Factor (TF) – Regulatory Element (TF Binding Site at/near Promoter/Enhancer) – Regulated Target Gene (TG) for 3 Brain Regions:

Full Alzheimer's Disease Gene Regulatory Network for the 3 Brain Regions				
Brain Region:	Total # of Unique Regulated Target Genes	Total # of Unique TFs	Total TF to Regulated Gene Relationships	Total TF, Regulatory Elements, and Target Gene Relationships
Dorsolateral Prefrontal Cortex (DLPFC)	13,511	670 *(out of 1,588 candidate Jaspar/Lambert TFs in DLPFC gene expression data)*	752,169	3,852,125
Hippocampus Ca1	11,972	351 *(out of 1,043 candidate Jaspar/Lambert TFs in Hippocampus gene expression data)*	169,292	2,810,102
Lateral Temporal Lobe (LTL)	13,791	402 *(out of 1,580 candidate Jaspar/Lambert TFs in*	65,321	161,404

		LTL gene expression data)		
Total (All 3 Regions Combined)	20,601	709	973,025	6,823,631

Table A.6: Metrics of Genes selected for Covid-19 severity prediction for Covid-19 positive human samples

Metrics of Genes Selected for Severity Prediction for Covid-19 Positive Human Samples						
Gene List	**Information**		**Original # of genes (Initial AD-Covid GRN)**	**Final AD-Covid GRN (Filtering for AD phenotype-related modules)**	**Optimal # of genes based on Recursive Feature Elimination with Stratified 5-Fold Cross Validation (CV) on Training Data**	**Final # of Genes in linear support vector classifier predictive model**
Dorsolateral Prefrontal Cortex (DLPFC)	AD-Covid GRNs	Brain Region-Specific	895	895	6	10
Hippocampus Ca1			1,305	1,146	7	10
Lateral Temporal Lobe (LTL)			670	322	16	10
Combined Regions		Overall	2,536	2,153	6	10
Published Genes (Benchmark)	Benchmark (Covid-19 Genes)		18 (filtering not applicable)		10	10

Then, we filtered our initial AD-Covid GRNs for each region for genes belonging to a "phenotype-enriched" gene co-expression module. The final Hippo. AD-Covid GRN genes belonged to any of 21 phenotype-enriched gene modules (shrinking genes from 1,305 to 1,146); the final LTL AD-Covid GRN genes were in any of 28 phenotype-enriched LTL modules (shrinking genes from 670 to 322).

Table A.7: Metrics of Predictive Models for Covid-19 severity prediction for Covid-19 positive human samples on training and testing data

Covid-19 Severity Prediction for Covid-19 Positive Human Samples from Median Normalized Patient Blood Gene Expression Data								
Gene List	**Information**		**Final # of Genes in linear support vector classifier predictive model**	**Training Data (80 samples)**			**Testing Data (20 samples)**	
				Stratified 5-Fold Cross Validation				
				Area Under the Curve (AUC)	**Accuracy (%)**	**Standard Deviation of Accuracy (%)**	**AUC**	**Accuracy (%)**
Dorsolateral Prefrontal Cortex (DLPFC)	AD-Covid GRNs	Brain Region-Specific	10	0.903	81.25%	0.0988	0.98	85%
Hippocampus Ca1			10	0.903	87.50%	0.10029	0.8	75%
Lateral Temporal Lobe (LTL)			10	0.972	85%	0.02096	0.87	80%
Combined Regions (All 3)		Overall	10	0.922	82.50%	0.05962	0.82	70%
Published Genes (Benchmark)	Benchmark (Covid-19 Genes)		10	0.863	77.50%	8.48%	0.79	60%

Table A.8: 36 AD-Covid genes

36 AD-Covid Genes					
ACTG1	BCL6	DUSP22	LILRA2	MYLIP	SF3B1
ANP32B	BOLA3	EMP3	LILRA6	PHF11	SMIM27
ANXA11	CD180	FCGR2A	LIMD2	PLEK	SPI1
AP1S2	CD82	GNG10	LYAR	RPS13	STAT5B
ARF5	COX7C	GPI	MPP1	RPS3	TM9SF4
ATM	DOK1	HCST	MTSS1	RPS7	TTC39C

§ A.4 Supplementary files and information

Please click the respective hyperlinks to navigate to the following publicly-available files for SNPheno.

File A1

Genes, modules, phenotypes, enrichments, TFs regulating gene modules: Hippocampus [💻🖱Link to data].

File A2

Genes, modules, phenotypes, and enrichments, TFs regulating gene modules: LTL [💻🖱Link to data].

File A3

Genes, modules, phenotypes, enrichments: DLPFC [💻🖱Link to data].

File A4

Filtered Gene Regulatory Network (GRN) for the Hippocampus CA1 (based on target genes associated with AD phenotypes or belonging to modules associated with AD phenotypes) [💻🖱Link to data].

File A5

Gene Regulatory Network (GRN) for the Lateral Temporal Lobe (LTL) [💻🖱Link to data].

File A6

22 shared AD and Covid-19 KEGG genes, 5 initial gene lists for initial AD-Covid GRNs (for Hippocampus CA1, LTL, DLPFC, All 3 combined, Published Covid-19 genes), final AD-Covid gene lists after filtering for genes in 'phenotype-enriched modules', Differential Expression Analysis genes found in AD-COVID GRNs for each region and median normalized gene expression for 100 samples for relevant input genes for Machine Learning analysis, testing samples used for Covid-19 prediction, final SVM linear kernel genes for 5 Covid severity models, predicted probabilities from each Covid severity SVM

model on training/testing data, Decision Curve Analysis (Net Benefit of each model for various probability thresholds) for Covid severity prediction on training/testing data [💻🖱Link to data].

File A7

Our 36 predicted AD-Covid genes used for AD Prediction in Superior Frontal Gyrus (SFG), AMP-AD genes used for AD Prediction, breakdown of SFG human cell-type samples used for models and the sample IDs used for testing, 8 sheets with raw gene expression values and binary dummy values (1 for each of 4 cell-types and for each of the 2 gene lists: AD-Covid or AMP-AD; AD-Covid model pools the 4 AD-Covid sheets for the training samples, and AMP-AD model pools the 4 AMP-AD sheets for the training samples), predicted probabilities from each logistic regression AD prediction model on testing data, Decision Curve Analysis for the AD predictive models on testing data [💻🖱Link to data].

File A8

Genome-Wide Association Studies (GWAS) SNPs used for Alzheimer's disease (AD) and/or severe Covid-19, expression quantitative trait loci (eQTL) data (slopes for brain and/or blood tissues), SNP-effected-GRN (AD and/or severe Covid-19 SNPs Interrupting Transcription Factor Binding Sites (TFBS) in Gene Regulatory Networks (GRNs) across any of the 3 Brain Regions, Metrics on SNPs impacting TFBSs (using 2 p-value cut-offs: our default GWAS p < 5e-5 ; stricter GWAS p < 5e-8). [💻🖱Link to data].

§ Chapter B: Supplemental Materials for NetREm

§ B.1 Supplementary methods and materials

Section §B.1.1: Mathematical Methods Work for NetREm:

§B.1.1.1 Integrating multimodal data & networks in NetREm workflow

Pre-processing gene expression data:

For the M cell samples, $X^{(0)} \in \mathbb{R}^{M \times N}$ matrix contains expression data for N predictor TFs, while $y^{(0)} \in \mathbb{R}^{M}$ is the expression vector for TG. $X^{(0)}$ dimensions are TG-specific based on N TFs. Expression data can be in different units (Zhao et al. 2021) and may undergo pre-processing using toolkits (e.g. Scanpy(Wolf et al. 2018), Seurat(Satija et. al 2015)). Nonetheless, we standardize $X^{(0)}$ to obtain X with $X_{ij} = \frac{X_{ij}^{(0)} - \overline{\mu_j^{(0)}}}{\sigma_j^{(0)}}$ where $\overline{\mu_j^{(0)}}$ is the mean and $\sigma_j^{(0)} > 0$ is standard deviation of TF_j in $X^{(0)}$. Each TF in X has $\overline{\mu_j} \approx 0$ and $\sigma_j \approx 1$ and any previous units are removed. Since r is invariant to scaling, original pairwise r are preserved. Similarly, we standardize $y^{(0)}$ to obtain y with $\mu_y \approx 0$, $\sigma_y \approx 1$. X and y are our final unitless, standardized expression input data.

§B.1.1.2 Step 1: Network regularized regression (Problem Definition)

Transforming NetREm to ElasticNet Problem:

Our objective function $f(c) = \frac{1}{2M}\left|\left|y - Xc\right|\right|^2 + \alpha\left|\left|c\right|\right|_1 + \beta c^T A c$ employs both L1 (Lasso) and L2 (Ridge) penalties to mitigate issues of collinearity. Even when the problem is transformed into a Lasso problem $\tilde{f}(c) = \frac{1}{2N}\left|\left|\tilde{y} - \tilde{X}c\right|\right|^2 + \alpha\left|\left|c\right|\right|_1$, the objective function's value $\tilde{f}(c)$ remains equivalent to $f(c)$ (subject to a constant), hence still addressing collinearity concerns. In the special case where $A = I$, the identity matrix (indicating a fully disconnected network prior), our objective transforms into the Elastic Net $f(c) = \frac{1}{2M}\left|\left|y - Xc\right|\right|^2 + \alpha\left|\left|c\right|\right|_1 + \beta\left|\left|c\right|\right|^2$. Given that we typically make our TG-specific input PPI network W fully-connected with artificial edge weight η for missing edges, we may set $w = \eta$ =1e-8 or some very small number so W is disjoint and reflects a scaled version of the Identity Matrix I.

§B.1.1.3 Step 2: Gene expression embeddings from network regression

Proof of Equivalence between Summation and Matrix Form:

The original network-based regularization term in the optimization problem is given by $\frac{\beta}{2}\sum_{i=1}^{N}\sum_{j=i}^{N} w_{ij}\left(\frac{c_i}{\sqrt{d_i}} - \frac{c_j}{\sqrt{d_j}}\right)^2 = \frac{\beta}{2}c^T Ac$. We show that this sum is equivalent to the form $\frac{\beta}{2}c^T Ac$ where the symmetric matrix $A \in \mathbb{R}^{N\times N}$ is defined as: $A = D^T(W\odot V)D$. Here, D is a diagonal matrix where $D_{ii} = 1/\sqrt{d_{ii}}$ and $W\odot V$ represents the element-wise (Hadamard) product of matrices W and V.

Expanding the Matrix Expression: We start by expanding the quadratic form: $c^T Ac = c^T D^T(W\odot V)Dc = (Dc)^T(W\odot V)(Dc)$. Please let $u = Dc$, such that $u_i = c_i/\sqrt{d_i}$. Thus, the expression becomes: $u^T(W\odot V)u = \sum_{i=1}^{N}\sum_{j=i}^{N} u_i(W\odot V)_{ij}u_j$. That is, $c^T Ac$ transforms to $\sum_{i=1}^{N}\sum_{j=i}^{N} u_i(W\odot V)_{ij}u_j$.

Examining $W\odot V$ Since matrix V is defined to modify the diagonal entries of W, we have $V = N\cdot I - 11^T$. Here, 1 is a vector of all ones. The Hadamard (element-wise) product $W\odot V$ adjusts W based on V, particularly affecting the diagonal elements. This matrix helps translates the interactions defined in W into the regularization context, taking into account the adjustments from V.

Proof of Equivalence: The goal is to show that $\sum_{i=1}^{N}\sum_{j=i}^{N} u_i(W\odot V)_{ij}u_j = 2\sum_{i=1}^{N}\sum_{j=i}^{N} w_{ij}(u_i - u_j)^2$

Since W is symmetric and without self-loops, and V modifies the diagonal entries to reflect subtracting the sum across rows, the off-diagonal terms of $W\odot V$ essentially remain w_{ij} for $i \neq j$, and: $(W\odot V)_{ij} = \begin{cases} d_i & \text{if } i = j \\ -w_{ij} & \text{if } i \neq j \end{cases}$.

Now, please note that the quadratic form expands as: $\sum_{i=1}^{N}\sum_{j=i}^{N} u_i(W\odot V)_{ij}u_j = \sum_{i=1}^{N} u_i^2 d_i - \sum_{i=1}^{N}\sum_{j\neq i} u_i w_{ij}u_j$, which, for $i < j$, can be rearranged into: $\sum_{i=1}^{N}\sum_{j=i}^{N} 2w_{ij}(u_i - u_j)^2$.

Conclusion: Hence, $c^T Ac$ accounts for each pair (i,j) with $i \neq j$ twice, as needed. Therefore, to match the original regularization term, we need the ½ factor: $\frac{\beta}{2}\sum_{i=1}^{N}\sum_{j=i}^{N} w_{ij}\left(\frac{c_i}{\sqrt{d_i}} - \frac{c_j}{\sqrt{d_j}}\right)^2 = \frac{\beta}{2}c^T Ac$.

Details regarding SVD on E matrix:

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 $\mathcal{S} = \{s_1, s_2, \dots, s_N\}$ 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, $\tilde{X}^T\tilde{X} = (N)(E)$. Hence, $\left(\sqrt{N}\Sigma^{\frac{1}{2}}U^T\right)^T\left(\sqrt{N}\Sigma^{\frac{1}{2}}U^T\right) = \tilde{X}^T\tilde{X}$. Thus, $\tilde{X} = \sqrt{N}\Sigma^{\frac{1}{2}}U^T$

where $\tilde{X} = \begin{bmatrix} | & | & \cdots & | \\ \widetilde{X_1} & \widetilde{X_2} & \ldots & \widetilde{X_N} \\ | & | & \cdots & | \end{bmatrix}_{N \times N}$; $\tilde{y} = \frac{\sqrt{N}}{M} \Sigma^{-\frac{1}{2}} U^T X^T y$. Here, $\Sigma^{\frac{1}{2}}$ and $\Sigma^{-\frac{1}{2}}$ are diagonal matrices with entries $\sqrt{\mathcal{S}}$ and $1/\sqrt{\mathcal{S}}$, respectively. The condition # of E with respect to a particular norm, $\kappa(E)$, measures how close E is to being singular. For the Euclidean norm, $\kappa(E) = \frac{s_{\max}}{s_{\min}}$. If $\kappa(E)$ is very large (e.g. 1e6), E may be nearly singular and ill-conditioned and results may be numerically unstable as inverting E (i.e. compute E^{-1}) can be difficult; in these unfortunate cases, $\geq$1 eigenvalue of E is $\approx$0, which is associated with increased variance of c^*. Small changes in inputs can cause large changes in c^*. In high-throughput data, there may be high multicollinearity, especially when $M \ll N$, so regression models may face problems related to a severely ill-conditioned Gram matrix $X^T X$ (Lê Cao et al. 2011; Hastie 2020). Adding βA to the Gram matrix to yield E may help eliminate potential issues arising from numerical instability in inverting $X^T X$, by ensuring E is well-conditioned (increasing smallest singular values of $X^T X$ to reduce $\kappa(E)$) and invertible. Nonetheless, our sparsity prior α on $\tilde{X}$ in $Eq(4)$, is an effective regularization penalty, helping tackle cases of an ill-conditioned E, stabilizing and guiding results, performing feature selection (Lê Cao et al. 2011; Hastie 2020; Kelner et al. 2021); by shrinking c^*, Lasso regression mitigates c^*'s sensitivity to errors/changes in inputs. Nonetheless, we endeavor to preempt any risk of an ill-conditioned E (i.e. $\kappa(E) \geq 1 \times 10^6$) via pseudo-inverses with an adaptive minimum threshold $t = \frac{s_{\max}}{1 \times 10^6}$. Singular value $s_i \in \mathcal{S}$ is problematic (corresponds to noise or redundant info) if $s_i \leq t$. In such cases, we set $\left[\Sigma^{\frac{1}{2}}\right]_i = 0$ (instead of $\sqrt{s_i}$) and $\left[\Sigma^{-\frac{1}{2}}\right]_i = 0$ (instead of $1/\sqrt{s_i}$). Our modification ignores directions (deemed unstable or insignificant in terms of data variance) like s_i, concentrating on strong singular values (most pertinent features), zeroing out roles of smaller ones in subsequent computations. This remedy enhances NetREm's interpretability, strengthening E's robustness against noise, addressing potential challenges related to M to N ratio (e.g. E being rank deficient).

Proof for the $A = D^T(W \odot V)D$ matrix:

A, D, W and $V \in \mathbb{R}^{N \times N}$ and all are symmetric matrices (i.e. $A = A^T, D = D^T, W = W^T, V = V^T$). $W \circ V$ multiplies $w_{ii} = \frac{\sum_{j=1,\ i \neq j}^{N} w_{ij}}{N-1}$ by $N-1$ to yield $d_i = \sum_{j=1,\ i \neq j}^{N} w_{ij}$ along the main diagonal and multiplies off-diagonal w_{ij} by -1 for $i \neq j$. Thus, $(W \odot V)_{ii} = d_i$ and $(W \odot V)_{ij} = -w_{ij}$ for $i \neq j$ where $(W \odot V) \in \mathbb{R}^{N \times N}$. Since D is symmetric, $D = D^T$. Since $D_{ii} = 1/\sqrt{d_i}$ and $D_{ij} = 0$ for $i \neq j$, then $D^T(W \odot V)$ will be:

$$\begin{bmatrix} 1/\sqrt{d_1} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{d_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sqrt{d_N} \end{bmatrix}_{N\times N} \begin{bmatrix} d_1 & -w_{12} & \cdots & -w_{1N} \\ -w_{21} & d_2 & \cdots & -w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ -w_{N1} & -w_{N2} & \cdots & d_N \end{bmatrix}_{N\times N} = \begin{bmatrix} \sqrt{d_1} & -w_{12}/\sqrt{d_1} & \cdots & -w_{1N}/\sqrt{d_1} \\ -w_{21}/\sqrt{d_2} & \sqrt{d_2} & \cdots & -w_{2N}/\sqrt{d_2} \\ \vdots & \vdots & \ddots & \vdots \\ -w_{N1}/\sqrt{d_N} & -w_{N2}/\sqrt{d_N} & \cdots & \sqrt{d_N} \end{bmatrix}_{N\times N}$$

. That is, $D^T(W \odot V) \in \mathbb{R}^{N\times N}$ where the main diagonal terms are $\sqrt{d_i}$ for $i = 1, \dots, N$ and the off-diagonal elements are scaled versions of the original weights w_{ij} divided by $\sqrt{d_i}$ based on the row i, for $i \neq j$. Hence, $A = D^T(W \odot V)D$ evaluates to:

$$\begin{bmatrix} \sqrt{d_1} & -w_{12}/\sqrt{d_1} & \cdots & -w_{1N}/\sqrt{d_1} \\ -w_{21}/\sqrt{d_2} & \sqrt{d_2} & \cdots & -w_{2N}/\sqrt{d_2} \\ \vdots & \vdots & \ddots & \vdots \\ -w_{N1}/\sqrt{d_N} & -w_{N2}/\sqrt{d_N} & \cdots & \sqrt{d_N} \end{bmatrix}_{N\times N} \begin{bmatrix} 1/\sqrt{d_1} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{d_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sqrt{d_N} \end{bmatrix}_{N\times N} = \begin{bmatrix} 1 & -w_{12}/\sqrt{d_1 d_2} & \cdots & -w_{1N}/\sqrt{d_1 d_N} \\ -w_{21}/\sqrt{d_1 d_2} & 1 & \cdots & -w_{2N}/\sqrt{d_2 d_N} \\ \vdots & \vdots & \ddots & \vdots \\ -w_{N1}/\sqrt{d_1 d_N} & -w_{N2}/\sqrt{d_2 d_N} & \cdots & 1 \end{bmatrix}_{N\times N}$$

. Thus, A has a main diagonal where $A_{ii} = 1$ for $i = 1, \dots, N$ and it has off-diagonal terms that are $A_{ij} = -w_{ij}/\sqrt{d_i d_j}$ where $i \neq j$.

The importance of βA for potentially reducing rank deficiency:

Here, $E = X^T X + \beta A$. Adding βA to yield E matrix effectively improves numerical stability. Rank-deficient matrices are often ill-conditioned because the presence of linear dependencies among rows or columns means that some singular values are ≈0, inflating $\kappa(E)$. The rank of a matrix signifies the maximum number of linearly independent columns or rows within the matrix. For instance, the maximum rank X can have is $\min(M, N)$. X is rank deficient if $\text{rank}(X) < \min(M, N)$, indicating there are redundant rows (if $N < M$, not all TFs are linearly independent) or columns (if $M > N$, then not all samples are linearly independent across TFs) in X, which do not add any new information or dimension to the space spanned by X.

$X^T X$ can have a maximum rank: $\min\big(\text{rank}(X), (\text{rank}(X^T X)\big) = \min(M, N)$. Since A behaves similarly to a graph Laplacian L, it has $N - 1$ linearly independent rows and columns (i.e. rank of $N - 1$ and rank deficient by 1). According to spectral graph theory properties(Klee and Stamps 2022), for a fully-connected graph, L has exactly one 0 eigenvalue (corresponding eigenvector is a constant vector of 1's representing the direction along which all TF node potentials are equal) and $N - 1$ positive eigenvalues. The second term (βA) has rank $N - 1$.

Theoretically, when $M < N - 1$, the first data term ($\frac{1}{M} X^T X$) in $Eq(3a)$, indeed is rank deficient with rank up to M since $\text{rank}(X^T X) = \min(M, N) = M$. And, $\text{rank}(A)$ is always $N - 1$ since it is a modified Laplacian matrix. When we add 2 semi-definite matrices (i.e. $X^T X$ and A), the rank is at least the minimum of the ranks of both matrices. That is, $\text{rank}(E) \geq \min\big(\text{rank}(X^T X), \text{rank}(A)\big) = \min\big(\min(M, N), \text{rank}(A)\big) = \min(\min(M, N), N - 1) = \min(M, N - 1) = M$. Hence, $\text{rank}(E) \geq M$. Then, the combined matrix terms in the embedding matrix $\frac{1}{N} \tilde{X}^T \tilde{X}$ (that is based on Singular Value Decomposition results of E) may have a rank that is at least M, showing how network

regularization may alleviate potential rank deficiency problems. In ideal cases with more samples than features ($N < M$), there is a reduced risk of overfitting and X may have full column rank of N; then the primary result of adding βA is a modified eigenstructure (i.e. structure and spectral properties) of $\frac{1}{M}X^TX$.In the case of full rank, we cannot guarantee that the condition number will be improved compared with X^TX, unless A is an identity matrix.

Proving Gene expression embedding for the target gene expression $\tilde{y}$:

Please note that here we verify Equation $3b$: $\frac{1}{N}\tilde{y}^T\tilde{X} = \frac{1}{M}y^TX$. Given that $\tilde{X} = \sqrt{N}\Sigma^{\frac{1}{2}}U^T$ and $\tilde{y} = \frac{\sqrt{N}}{M}\Sigma^{-\frac{1}{2}}U^TX^Ty$.

Then, $\tilde{y}^T = \frac{\sqrt{N}}{M}\left(\Sigma^{-\frac{1}{2}}U^TX^Ty\right)^T$. This expands to $\tilde{y}^T = \frac{\sqrt{N}}{M}\left(\Sigma^{-\frac{1}{2}}U^TX^Ty\right)^T = \frac{\sqrt{N}}{M}(y)^T(X^T)^T(U^T)^T\left(\Sigma^{-\frac{1}{2}}\right)^T =$

$\frac{\sqrt{N}}{M}y^TXU\left(\Sigma^{-\frac{1}{2}}\right)^T$. Here, $\Sigma \in \mathbb{R}^{N\times N}$ is a diagonal matrix of the singular values $\{s_1, s_2, \dots, s_N\}$ of E where $\Sigma^{-\frac{1}{2}} =$

$$\begin{bmatrix} s_1^{-1/2} & 0 & 0 & \cdots & 0 \\ 0 & s_2^{-1/2} & 0 & \cdots & 0 \\ 0 & 0 & s_3^{-1/2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & s_N^{-1/2} \end{bmatrix}_{N\times N} = \begin{bmatrix} 1/\sqrt{s_1} & 0 & 0 & \cdots & 0 \\ 0 & 1/\sqrt{s_2} & 0 & \cdots & 0 \\ 0 & 0 & 1/\sqrt{s_3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1/\sqrt{s_N} \end{bmatrix}_{N\times N}.$$ Hence, $\left(\Sigma^{-\frac{1}{2}}\right)^T = \Sigma^{-\frac{1}{2}}$.

Then, $\tilde{y}^T = \frac{\sqrt{N}}{M}\left(\Sigma^{-\frac{1}{2}}U^TX^Ty\right)^T = \frac{\sqrt{N}}{M}(y)^T(X^T)^T(U^T)^T\left(\Sigma^{-\frac{1}{2}}\right)^T = \frac{\sqrt{N}}{M}y^TXU\,\Sigma^{-\frac{1}{2}} \rightarrow$ Thus, $\tilde{y} = (\tilde{y}^T)^T$. Based on this, $\tilde{y} = \left(\frac{\sqrt{N}}{M}y^TXU\,\Sigma^{-\frac{1}{2}}\right)^T = \frac{\sqrt{N}}{M}\left(y^TXU\,\Sigma^{-\frac{1}{2}}\right)^T \rightarrow \tilde{y} = \frac{\sqrt{N}}{M}\left(\Sigma^{-\frac{1}{2}}\right)^T U^TX^T(y^T)^T$. Since $\beta > 0$ is imposed, it is mandatory to have a PPIN among TFs so $\tilde{X}$ encodes not only principal components of X but also the PPIN of relations among TFs. 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 info.

§B.1.1.4 Output 1: Identification of novel cell-type TFs for regulating TG

Metrics to evaluate the regression performance of Target Gene (TG)

For each TG in a cell-type (and context), we evaluate the performance of NetREm's regression model c^* (used to define our TF-TG regulatory links) in train and test expression data. That is, we compare predicted values $\widehat{\tilde{y}} \in \mathbb{R}^{N\times 1}$ to actual embeddings $\tilde{y}$ by relating signal (ground truth $\tilde{y}$) to noise (errors between $\widehat{\tilde{y}}$ and $\tilde{y}$) in training and testing gene expression data:

- Mean Square Error (averages the squared differences between predicted and actual expression values for TG): $MSE = \frac{1}{N}\sum_{v=1}^{N}\left(\widehat{\tilde{y}_v} - \tilde{y}\right)^2 = P_{noise}$
- normalized MSE (helps understand quality or purity of a signal): $NMSE = \frac{P_{noise}}{P_{signal}}$.

- signal-to-noise ratio (measures signal strength relative to background noise): $SNR = 10 \times \log_{10}\left(\frac{P_{signal}}{P_{noise}}\right)$
- peak SNR ($PSNR = 10\log_{10}\left(\frac{MAX_I^2}{P_{noise}}\right)$).

Here, $P_{signal} = \frac{1}{N}\sum_{v=1}^{N}\tilde{y}^2$ is the mean squared value of $\tilde{y}$ and MAX_I^2 is max possible signal intensity (default: $\max|\tilde{y}|$). Generally, ideal regression solutions have low MSE and NMSE values and high SNR and PSNR values. NetREm can unearth novel, cell-type-specific TFs involved in TG regulation. These N^* regulatory links likely capture truer biological interactions among TFs.

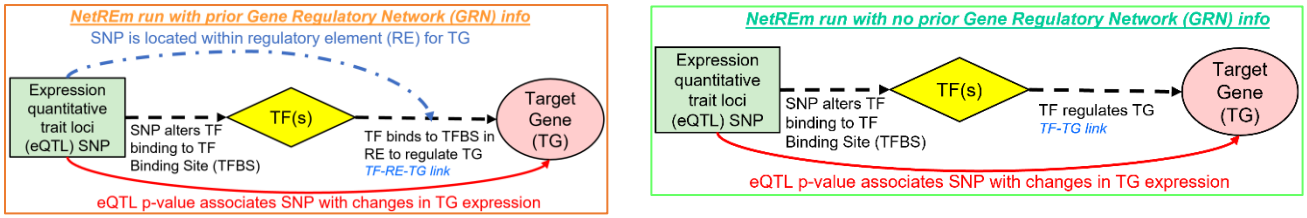
Incorporating Prior Gene Regulatory Networks (GRNs) to derive TF-RE-TG regulatory networks

We identify regulatory elements (RE) for a given TG. Then, we determine TFs likely to bind directly to or associate indirectly with TF Binding Sites (TFBSs) on these REs. We input N TG-specific candidate TFs to NetREm for TG. Then, we overlay NetREm's N^* TF-TG regulatory links for TG with this prior GRN (initial TF–RE –TG links for N TFs for TG). Then, we can uncover the highly confident final TF-RE-TG links for our final set of N^* TFs for TG. These final TF-RE-TG correspond to the N^* TFs that are selected for regulating the expression of the TG. These TFs may bind directly or indirectly associate with these REs. Thus, the prior GRN links can help annotate our final regulatory links. Please note that this is subjective and is based on steps we detail to obtain a prior GRN.

Mapping SNPs onto NetREm's TF-TG or TF-RE-TG regulatory networks

To annotate/validate these inferred relationships, we may map Single Nucleotide Polymorphisms (SNPs) on top of these networks. We use motifbreakR (Coetzee et al. 2015), a tool designed to predict the impact of SNPs on TF binding to their binding sites (TFBSs). We use not only strong effects but also weak effects for motifbreakR. We primarily use strong effects for TFs that are predicted to directly associate to the Regulatory Element (RE) to regulate the TG (i.e. TFs from the motif-based GRN that are retained in the pruned TF list); for simplicity, we do not consider the TFs that may indirectly associate with these REs. We use PWMs from motifbreakR as well as our own comprehensive PWMs. We use all 3 methods "default", "ic", "log" for computing SNP-TF effects.In this way, we may predict pathways from non-coding SNPs to disease phenotypes, via TFs with altered TF binding to TFBSs, impacting TF interactions with each other and ultimately leading to changes in expression of the TG.

- alleleDiff > 0: SNP may increase TF binding (boosts its affinity) to its TFBS on the RE.
- alleleDiff < 0: SNP may decrease TF binding (disrupt its motif recognition and binding) to TFBS.



We then map these SNP-TF links on top of these along with the expression quantitative trait loci (eQTL: eSNP-eTG links) links on these networks. In eQTLs, we have eSNP-eTG that means this expression quantitative trait SNP (eSNP) is linked to expression quantitative trait changes in TG (eTG).

Left figure: If we use prior (reference) GRN knowledge info, then we may have final TF-RE-TG links from NetREm. We check that the SNP impacts TF binding (either increasing or decreasing TF binding), that this TF binds to a regulatory element (RE) for the TG to regulate the TG (based on final TF-RE-TG links), this SNP falls within this same RE, this SNP associates with changes in TG expression (based on eQTL links).

Right figure: If we do not use any prior GRN knowledge info, then we have final TF-TG links from NetREm. We check that the SNP impacts TF binding (either increasing or decreasing TF binding), that this TF regulates TG (based on final TF-TG links), this SNP associates with changes in TG expression (based on eQTL links). These networks are not as confident as those from the left figure given the criteria is not as strong as it is for the left figure.

Section §B.1.2: Benchmarking for NetREm with No Prior GRN Knowledge

We evaluate how effectively NetREm predicts meaningful TF-TG and TF-TF coordination networks in real-world cases, even without any prior GRN info (i.e. all TGs have same N candidate TFs). For each TG, $N = \mathcal{N}$, where $\mathcal{N}$ is global set of TFs; TGs that are TFs have $N = \mathcal{N} - 1$ TFs. We anticipate our inferred TF-TG regulatory and TF-TF coordination networks in our upcoming SC and AD applications (apps) will be improved, given that we incorporate prior GRNs. We use multiomics data to derive initial TF-RE-TG links; these prior GRNs perform initial feature selection, resulting in a small, customized list of N promising and biologically-relevant TFs for a given TG as input to NetREm. N may vary across TGs.

NetREm enhances robustness through PPIN regularization, prioritizing cell-type-specific TF-TF coordination. This is advantageous in high-dimensional $N > M$ data, where overfitting and multicollinearity can lead to spurious r among variables (Campos et al. 2019; Hoefsloot et al. 2008). Single-cell expression data, prone to high sparsity (≈70% of entries ≈ 0), often fails to accurately represent expression distributions (Nguyen et al. 2020).

Further, expression data often contains noise of both direct and indirect TF-TG interactions (Escorcia-Rodríguez et al. 2023). The inherent complexity of gene regulation in eukaryotes (e.g. humans, mice) can also obscure TF-TF correlations, complicating the identification of functionally coordinated TFs (Nie et al. 2011). TFs co-regulating common TGs typically exhibit high r due to similar expression profiles, suggesting potential co-association and cooperativity (Roy et. al 2020; Nie et al. 2011; Wang et al. 2016). This r can be problematic when predicting TF-TG regulatory links using expression data, as many correlated TFs may be causally related to TG expression (Ahsendorf et al. 2017; Ouyang et al. 2009). Even uncorrelated TF_i-TG r can indicate broader coordination among many TFs including TF_i (Zaborowski and Walther 2020). BRMs struggle with highly correlated features, often selecting independent TFs or omitting true co-regulating TFs, compromising the integrity of their TF-TG regulatory network (Roy et. al 2020; Nicodemus and Malley 2009). Our optimization incorporates PPIN structures among TFs to encourage similarity among strongly connected TFs, thereby accounting for noise due to TF coordination (Parab et al. 2022). NetREm discerns and assigns each TF's influences with superior generalizability and consistency, capturing intricate interactions that may be oversimplified by BRMs. NetREm's balanced approach to predictive accuracy and structural interpretability makes it a computationally sound method. This ability to incorporate comprehensive prior data, such as PPINs, into GRN predictions from expression data is a significant advantage of NetREm over traditional methods. This integration not only enhances the accuracy of the GRN but also ensures that essential but less prominent TFs are recognized in the regulatory network. Such an approach is vital for understanding the full spectrum of TF activities in specific cell types, demonstrating the importance of using extensive prior info in GRN predictions, a capability implemented by NetREm (Dibaeinia and Sinha 2020).

§B.2.1: Cell-type-specific TF-TG regulatory networks (output 1)

To evaluate our networks holistically, we compare NetREm with 4 BRMs in terms of predicted signed (c^*: +/-) TF-TG regulatory links. For this, we use SERGIO (Dibaeinia and Sinha 2020) to generate 6 realistic datasets ($M = 70$ or 700 train; noise %: 30, 60, 90) for 1,250 TGs and $\mathcal{N}$ = 207 TFs based on a signed ground truth atlas GRN in hESCs. These 6 datasets enable us to evaluate performance based on M-to-N ratio and noise that may obscure true signals. These models (except GRNBoost2) use c^* to assign TF roles (activator, repressor); NetREm uses c^* to also infer the nature (cooperative, antagonistic) of coordination among TFs for regulating TGs. Overall, NetREm achieves the highest precision in predicting signed-TF-TG links, suggesting its stronger reliability in identifying true TF-TG links, assigning accurate TF roles, prioritizing promising links with a reduced risk of False Positives (FPs)

(**Figure B.6**). When fit with LassoCV, NetREm's TF-TG accuracy is superior to that of benchmarks (GRNBoost2, BRMs), underscoring its robustness to noise. In fact, GRNBoost2 has the lowest accuracy even with complete data usage (M = 100 or 1,000).

NetREm enhances robustness through PPIN regularization. To see this, we benchmark NetREm on real-world single-cell gene expression data in humans (hematopoietic stem cells: HSCs) and mice (mESCs, mDCs) (**Fig. B.7-B.8**, **Tables B.4-B.6**). These datasets are noisy, sparse, high-dimensional ($N \gg M$) and reflect the inherent complexity of eukaryotic gene regulation. We evaluate our TF-TG links against respective gold standard GRN TF-TG links that have no c^* sign info (McCalla et al. 2023; Zhang et al. 2023). Our findings corroborate that NetREm has higher sensitivity (as it incorporates biological info (Shojaie and Michailidis 2009)) for identifying relevant biomarkers, but has lower specificity compared to Lasso and ElasticNet (Li and Li 2008). Tuning β and/or α can adjust NetREm's results; for instance, increasing α may reduce sensitivity and increase specificity.

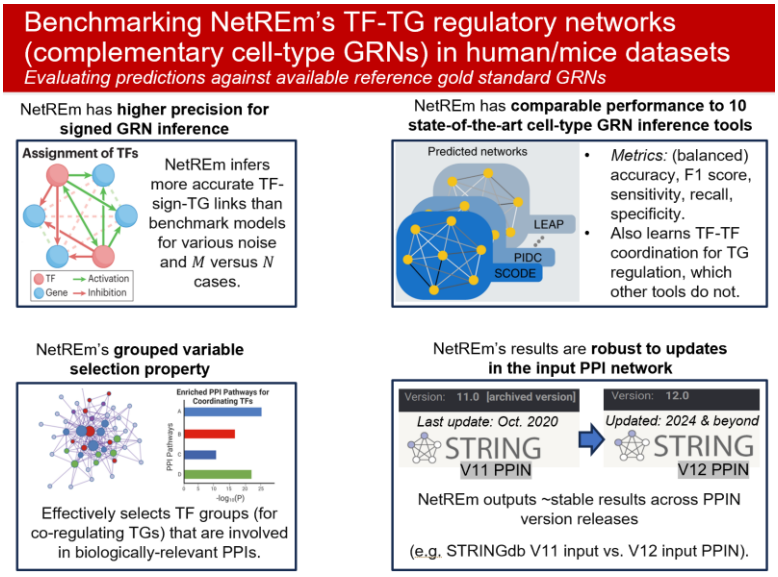
Without the use of prior PPI info that NetREm incorporates, predicting TF-TG links using single-cell expression data alone can be problematic. This data often fails to accurately represent expression distributions (Nguyen et al. 2020). It is further challenging to identify TFs that truly coordinate solely based on TF-TF correlations in this data (Campos et al. 2019; Hoefsloot et al. 2008). Correlated TFs with similar expression profiles may either coordinate to causally co-regulate TGs (Ahsendorf et al. 2017; Ouyang et al. 2009; Wang et al. 2016) or merely be spuriously co-associated. Linear Regression and Ridge tend to retain correlated TFs, Lasso often drops them (potentially omitting co-regulating TFs), ElasticNet balances Ridge and Lasso. Thus, BRMs may struggle to identify functionally coordinated TFs for TG regulation (Nie et al. 2011), compromising the integrity of their TF-TG regulatory networks (Roy et. al 2020; Nicodemus and Malley 2009). On the other hand, NetREm considers not only TF-TF correlations but also info on TF-TF PPIs, and helps account for noise in expression data due to TF-TF coordination (Parab et al. 2022). NetREm can thus uncover TF-TF interactions for TG regulation, discerning each TF's influence with superior generalizability and consistency, capturing intricate relations that BRMs cannot.

NetREm's optimization incorporates PPIN structures among TFs to encourage the selection of strongly grouped TFs involved in known, biological PPIs. We highlight this grouped variable selection property (Li and Li 2008) in HSCs where we run NetREm and benchmarks for 10,588 TGs and $\mathcal{N}$ = 178 TFs (**Table B.4**). ElasticNetCV and LassoCV are limited in identifying TF-TG links for TGs. For instance, both notably miss regulation of *ATF2*, a pivotal TF in stem cells (Ju et al. 2023). On the other end, Linear Regression and RidgeCV

predict that *ATF2* is regulated by 177 TFs, illustrating their alarming potential for FPs. However, for $\beta = 10$, NetREm uncovers 8 final TFs for regulating *ATF2* where all but WHSC1 are substantiated by gold standards (Zhang et al. 2023). We note the same for *BRD2, RNF167, DUSP2*. NetREm flags groups of verified and novel coordinating TFs connected along biologically meaningful, cell-type PPIs (Li and Li 2008) for these 4 TGs (**Fig. B.7**). For instance, TFAP4, 1 of our 10 novel TFs for *RNF167*, is involved in adipogenesis and negative regulation of cell population proliferation with 3 of 17 substantiated final TFs. These examples not only validate NetREm's applicability to complex biological systems but also show, directly, its ability to effectively identify accurate TF groupings from a large pool of candidate TFs.

NetREm's results are robust to changes in the input PPIN from old to new releases (**Table B.5**). To demonstrate this, we run NetREm($\beta = 1$, $\alpha = 0.05$) and compare results in mESCs for old STRINGdb version 11 (V11) versus updated V12 PPINs. Indeed, the mouse PPIN evolves from V11 to V12. For instance, 25 singleton TFs (of 195 mESC TFs) in V11 have known PPIs in V12. Nonetheless, our evaluation metrics results are similar.

NetREm's TF-TG regulatory networks are complementary to GRN tools. To show this, we apply NetREm to expression data in mDCs that has been used to benchmark 10 SOTA cell-type GRNs (McCalla et al. 2023). NetREm has comparable performance with SOTA GRNs across various metrics (sensitivity, specificity, F1 Score, balanced accuracy, overall accuracy) (**Table B.6, Fig. B.8**). No method outperforms another for predicting TF-TG links. NetREm, however, infers TG-specific B and cell-type-specific TF-TF coordination networks $\bar{B}$, which the others cannot do. We summarize our core benchmarking results for NetREm's TF-TG regulatory networks:

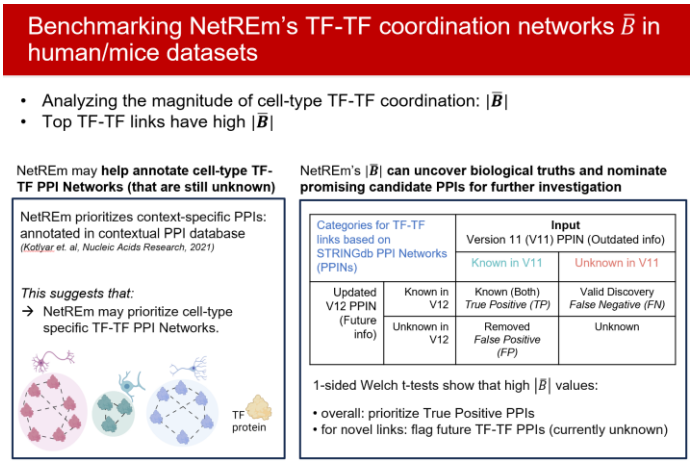


§B.2.2: Cell-type-specific TF-TF coordination scores $\bar{B}$ (output 2)

We adopt innovative approaches to benchmark NetREm's $\bar{B}$ for TF-TF pairs for TG regulation. Specifically, we use Welch 1-sided tests (p-adj < 0.05) to analyze its performance with public STRINGdb version 11 (V11) as the input PPIN, a proxy for outdated info. V12 represents future insights. TF-TF pairs are categorized into 4 sub-groups: those retained in both PPINs (TPs), those removed in V12 (FPs), those absent in V11 but present in V12 (FNs), and those absent in both versions. NetREm's $|\bar{B}|$ values across these sub-groups are analyzed to see if high $|\bar{B}|$ values for novel TF-TF links (w = 0.01) indicate valid future discoveries (FNs) and if NetREm effectively prioritizes TPs over FPs (results for mESCs, mDCs, PBMCs: **Tables B.7-12, Figures B.9-11)**. For instance, in mESCs, $|\bar{B}|$ is significantly higher for V11 than non-V11 links and for TPs than for FNs, indicating that high $|\bar{B}|$ often reflects known PPIs. $|\bar{B}|$ is also higher for FNs than for FPs, showing NetREm's ability to flag links for removal. In terms of novel links (not known in V11 but artificially added with weight $w = 0.01$), $|\bar{B}|$ is higher for FNs than for those whose status remains unknown, suggesting that $|\bar{B}|$ can uncover biological truths and effectively nominate promising candidate PPIs for further investigation. Thus, NetREm prioritizes known TPs and flags actual future TF-TF PPIs that are yet unknown.

Further, we benchmark NetREm's performance against that of a contemporary tool, RTNduals (Chagas et al. 2019). Both tools recognize that each TG in a GRN may link to many regulators based on direct and/or indirect interactions among TFs. We compare NetREm run with human V11 PPIN against RTNduals, using the same expression data and TFs for 13 human contexts where RTNduals returns its k final TF-TF links (as it may fail to yield results for some data) (**Figure B.12**). We select our top k links (high $|\bar{B}|$) and evaluate both using known PPIs from V12 and other sources (Göös et al. 2022). Except for Microglia (Mic: resident macrophage immune cell in CNS) (Lake et. al 2018), NetREm has fewer poor results (FPs and/or unknown) for 12 contexts, learning a greater % of actual PPIs; this further underscores NetREm's unparalleled predictive prowess in prioritizing TPs and in leveraging historical PPIs to forecast and discover actual PPIs not yet known in the existing PPIN. This is encouraging as PPINs are largely incomplete: a small fraction of ≈130-650k potential human PPIs are identified in experiments (Sevimoglu and Arga 2014; Venkatesan et al. 2009) that may include FPs (Yu et al. 2020). A strength of NetREm is that it not only weights known PPIs, but also estimates coordination when it may not be possible to observe PPIs.

Moreover, we gauge whether NetREm can help contribute to the detection of cell-type-specific TF-TF PPIs. This is an on-going research problem as existing PPINs are generally global and are not cell-type-specific. We use a Contextual PPI database (CPPID) (Kotlyar et. al 2022), that enriches PPI networks by providing annotations for over 243 context-specific terms, although it lacks cell-type specificity; it labels known PPIs with these context terms. We use this CPPID to see if our top coordination links (i.e. $|\bar{B}| \geq 85$) across contexts reflect context-specific annotations (i.e. are enriched with context-specific terms). Despite this CPPID not being cell-type-specific, it still helps show that top links in Mic and pooled SCs (Eraslan et. al 2022), are biologically relevant, enriched for nervous system (NS)-related terms (**Figure B.13**). Similarly, strong TF-TF links in immune Peripheral Blood Mononuclear Cells (PBMCs) associate with immune-related terms. Currently, there is no direct method to validate that these top links are specific to particular cell types, posing an ongoing research challenge. However, we can tentatively extrapolate that if NetREm's links are context-specific, they might also be indicative of cell-type specificity. This assumption positions NetREm as a potential pioneer in identifying cell-type-specific TF-TF PPIs, exploring new frontiers in the field.

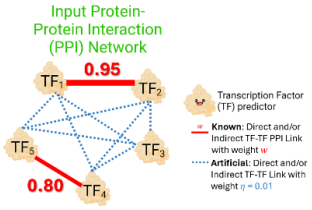


§B.1.2.3: SCENIC applied to human myelinating (mSCs) and non-myelinating (nmSCs) SCs in Dorsal Root Ganglion (DRG):

While state-of-the-art (SOTA) tools like SCENIC (Aibar et al. 2017) may provide indirect insights into TF-TF interactions (e.g. TFs co-regulating many common TGs), they sadly focus on TFs with strong binding and do not incorporate other prior info. To demonstrate, we perform comparative analysis of SCENIC in SCs. After running GRNBoost2 with 1,839 TFs to detect gene modules showing co-expression alongside TFs, we find 16,109 and 14,482 TGs for 1,572 and 1,432 TFs (including core SC TF TEAD1) in nmSCs and mSCs. SCENIC's cis-regulatory motif analysis retains modules with significant enrichment for the appropriate upstream regulator, eliminating

indirect TGs lacking motif support. SCENIC detects 640 TFs and 9,241 TGs in nmSCs and 522 TFs and 9,622 TGs in mSCs but fails to identify TEAD1 in both SCs, detecting 3 TEAD family TFs instead. By integrating PPIs, NetREm detects true GRN relations for core TFs with relatively weak binding info (e.g. TEAD1) whose signals may be drowned out. Hence, it is important to integrate additional omics data (Dibaeinia and Sinha 2020), like PPINs as prior info, into GRN prediction from expression regression, which is precisely what NetREm accomplishes.

Section §B.1.3: Simulation Study:



Utilizing prior knowledge of networks, NetREm can identify and allocate influences of each TF, providing reliable and nuanced insights into individual contributions, despite the presence of existing intercorrelations. We illustrate this capability by running 2 larger simulation analyses in **§B.1.3.1** and **§B.1.3.2**, each involving 1,000 simulations for the same underlying data. In each simulation, the random seed is changed. Relative pairwise correlations among TFs are nearly identical between training and testing data. For each of the two simulations below, we maintain the original simulated prior protein-protein interaction (PPI) network ($TF_1 - TF_2$ edge weight of 0.8, $TF_4 - TF_5$ edge weight of 0.95, default edge weights of 0.1) and $cor(TF, TG)$ = [0.9, 0.5, 0.4, -0.3, -0.8], as in **Figure 3.2A**. We do not fit any y-intercept. Across simulations, our NetREm models continue to standardize X and standardize y. The results of these additional simulation analyses underscore the necessity for careful tuning and adjustments of α and β to help NetREm achieve optimal complexity and dependable predictiveness.

§B.1.3.1 Testing Robustness

This analysis tests the robustness of NetREm (measured by the standard deviation) compared with that of the 4 benchmark regression models (BRMs). All 4 BRMs are fit with defaults and no y-intercept. We run benchmarks and NetREm (with LassoCV) on 1,000 simulations of the respective random generated data (changing random seed),]; this setup allows for variations in pairwise correlations among TFs in each simulation. Please note that for each pairwise Welch test, we use the R software t.test function default of var.equal = False. We compare the coefficients predicted by NetREm for each of the 5 predictors with those predicted by each of the 4 respective BRMs (RidgeCV, Linear Regression, ElasticNetCV, LassoCV) where we test the following alternative hypotheses (H_A) based on the sample means $\widehat{\mu_{TF}}$ for each respective TF predictor (assuming that the Null Hypothesis H_o asserts that there is NO statistically significant difference in sample mean values for the coefficients for that given TF between NetREm and that benchmark model): For TF_1, TF_2, TF_3: H_A: $\widehat{\mu_{Netrem,\ TF_i}} > \widehat{\mu_{Benchmark,\ TF_i}}$; For TF_4, TF_5: $\widehat{\mu_{Netrem,\ TF_i}} <$

$\widehat{\mu_{Benchmark,\ TF_t}}$. We guide these 1-sided t-tests (at the 5% level: p-value < 0.05 to reject H_o for statistical significance) based on what we strive to illustrate in this demo (i.e. indicated by change in signs < or > for H_A). Our results (**Table B.1**) are statistically significant for TFs 2 to 4 compared to BRMs. Comparative robustness checks across 1,000 simulations show NetREm has less variability and more stability for all c^* (**Table B.2, Figure B.3A**).

§B.1.3.2 Benchmarking β

The 2nd analysis focuses on the impact of network-constrained prior hyperparameter β on NetREm coefficients, standard deviation, and test performance (measured by the mean square error (MSE)). Here, we benchmark β by re-running 1,000 simulations to investigate the impacts of altering β (across 13 test values: 0.1, 0.25, 0.5, 0.75, 1, 2.5, 5, 7.5, 10, 25, 50, 75, 100) on NetREm's performance, holding all else constant; thus, we fix α at α^* and use Lasso; we compare NetREm models with one another based on varying the network constrained hyperparameter β and holding the lasso hyperparameter term α constant at 0.1. This procedure elucidates NetREm's adeptness at managing variance (e.g. std) by incorporating a manageable level of bias, which is key for optimizing the bias-variance tradeoff, preventing overfitting, enhancing generalization to unseen data; nonetheless, there is a critical point beyond which excessive network regularization penalty can be detrimental, leading to models overly constrained by the predefined network structure.

Bias (e.g. test MSE) increases monotonically as β increases. Our network regularization term scaled by β can aid with managing variance (e.g. standard deviation (std) of coefficients based on changes in data) of our model, even if it introduces some level of bias. This key adjustment helps models generalize better to unseen data by preventing overfitting, a situation often characterized by low bias but high variance. This procedure elucidates NetREm's adeptness at managing variance by incorporating a manageable level of bias, which is key for optimizing the bias-variance tradeoff, preventing overfitting, enhancing generalization to unseen data. Coefficients for TF_2 (0.101 to 0.283) and TF_4 (-0.037 to –0.107) increase in magnitude and standard deviation of all 5 coefficients decrease as β increases to 5. In any case, β = 7.5 is still alright in terms of standard deviation values. However, there is a critical point beyond which excessive network regularization penalty can be detrimental, leading to models overly constrained by the predefined network structure. Elevated β values, specifically at 50, 75, and 100, cause NetREm to struggle as it strictly adheres to network relations, leading to increased bias and variance. This deteriorates not only predictive performance but also variance, observed by suboptimal coefficients, increased std. Thus, meticulously tuning both α and β can secure a balanced model complexity and maintain robust, reliable

predictive performance. 1,000 simulations for 13 β values (**Table B.3, Figure B.3B-C**), illustrate bias-variance trade-offs where excess β over-constrains models, causing NetREm to struggle and predictions to suffer.

§B.3.3 Sparsity selection for main simulation study

We experiment with altering various levels of sparsity of our underlying gene expression data (e.g. 10%, 50%, 60%, 70%) and rerunning **Figure 3.2** analysis with the new data. Please note our results are in **Figure B.2**. Results are consistent for different sparsity levels (e.g. 10, 50, 60, 70%) of expression data.

§B.3.4 Simulation with various sparsity levels:

To tackle the issue of gene expression sparsity, notably pertinent to TF expression in single-cell analyses, we performed a simulation study with $M = 100$ and $N = 500$ across different sparsity levels (0.70, 0.85, 0.90). In this context, 70%, 85%, and 90% of the gene expressions are 0s. We evaluate NetREm's performance using mean squared error (MSE) and note an improvement in performance with increasing sparsity levels. This table below affirms NetREm's adeptness at managing data sparsity effectively:

Sparsity	0.70	0.85	0.90
MSE	0.0239	0.0183	0.00903

§B.3.5 Testing various M versus N cases (no ill-conditioned results)

For each scenario (i.e. $N \ll M, N = M, N \gg M$), NetREm's resulting E has SVD singular values $\mathcal{S}$ in a small range so E is not ill-conditioned: condition # $\kappa(E) = \frac{s_{\max}}{s_{\min}} \ll$ 1e6 (**Fig. B.4**). If there are any potentially ill-conditioned results (**§B.3.7**), our truncated SVD approach will help ensure numerical stability. We do use Lasso regression, which will also assist with helping increase robustness of results.

§B.3.6 Adapting Simulation Study for $N > M$ case

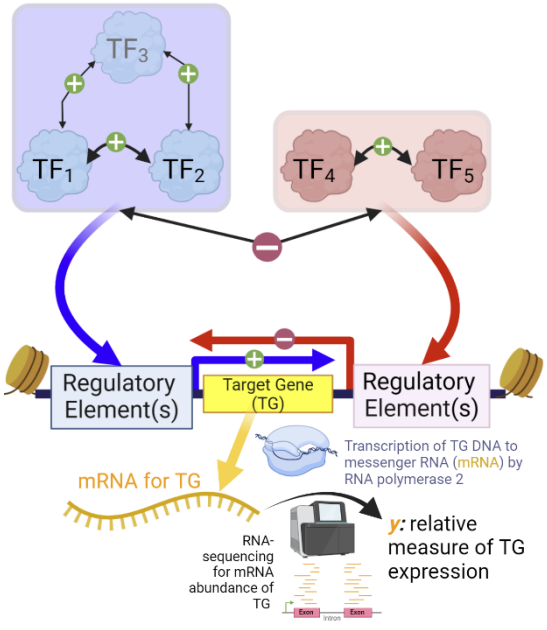
We adapt **Fig. 3.2A**, for a $N = 6 > M = 5$ case with 40% sparsity (**Fig. B.5A**). We design this envisioning TF_6 is not a final TF. TF_1 to TF_6 expression r with y are ≈ [0.85, 0.34, 0.13, -0.28, -0.88, -0.04] (**Fig. B.5B**). We use **Fig. 3.2A** PPIN where TF_3 and TF_6 nodes are added artificially. Internally, NetREm creates E (**Fig. B.5C**) summing expression- $(\frac{1}{M}X^TX)$ and PPIN- (βA) -based matrices (**Fig. B.5D**). Here, $\text{rank}(X^TX) = 4$ is deficient as it is not $\min(N, M) = 5$; $\text{rank}(\beta A) = N - 1 = 5$. NetREm may help improve rank deficiency, especially for $N > M$: E has full rank of 6 = N and is well-conditioned, so results are stable (**Fig. S5E**). Only NetREm accurately learns TF_2's $c^* > 0$ (**Fig. B.5F**). While LassoCV and ElasticNetCV also remove orphaned TF_3 and TF_5, they also eliminate TF_4 and have higher MSE (**Fig. B.5G**). Sensitivity analysis (**Fig. B.5H**) is consistent with **Fig. B.2D-E**. Activators TF_1 and TF_2 cooperate, repressors TF_4 and TF_5 cooperate; other interactions among the N^*= 4 final TFs are antagonistic (**Fig. B.5I**).

§B.3.7 Simulation with various M and N

We enhance stability by setting small singular values (e.g., $\sigma_i < 10^{-6}\sigma_{max}$) to zero, resulting a truncated Σ_{trunc}. Similarly, we adjust the inverse ${\Sigma_{trunc}}^{-1}$ so that the inverse elements corresponding to small singular values are set to zero. To illustrate the necessity of this approach, we simulated various cases where $M < N, M = N$, and $M > N$, computing the conditional numbers of $E = \frac{X^T X}{M} + \beta A$ for various values of β as shown in the table below:

β	M	N	Conditional Number $\kappa(E)$
10^{-6}	100	500	1.39×10^{7}
10^{-6}	300	300	6.28×10^{6}
10^{-6}	500	100	3.94×10^{6}
1	100	500	1.83
1	300	300	8.84
1	500	100	7.32
10^{6}	100	500	1.14×10^{6}
10^{6}	300	300	1.56×10^{6}
10^{6}	500	100	1.69×10^{6}

This method demonstrates how conditional numbers $\kappa(E)$ vary across different configurations, revealing that the condition number can be significantly large ($> 10^6$), thus showing the necessity of employing truncated SVD. The analysis also shows that when β is small, $E = \frac{X^T X}{M} + \beta A$ primarily depends on the data term $\frac{X^T X}{M}$, where the conditional number decreases as the number of measurements increases (from $M < N$ to $M > N$). When β is large, $\kappa(E)$ mainly depends on the network prior term βA which is determined by the PPIN and is independent of M.

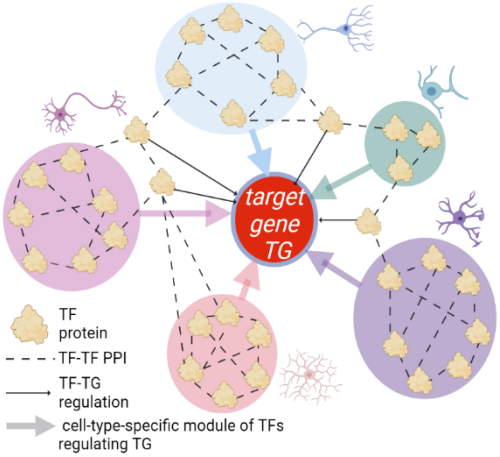


§B.3.8 Biological Interpretation of Main Simulation

Activators: positive NetREm coefficient $c^* > 0$. Repressors: negative NetREm coefficient $c^* < 0$. Please note that this figure on the left shows how we may interpret NetREm's main results from **Figure 3.2.** It is an illustration of cell-type-specific TF-TF coordination interactions that are involved in gene regulation of TG with relative gene expression y. The TF-TG regulatory network shows TF_1, TF_2, and TF_3 are activators of TG (work to increase TG expression y) since they have positive coefficients: $c^* > 0$. Given their negative coefficients for regulating TG (i.e. $c^* < 0$), we assign

TF_4 and TF_5 as repressors of TG (work to decrease y). We utilize both outputs of NetREm: cell-type TF-TG Regulatory Network (given by coefficients c^* and the TG-specific TF-TF interactions (given by B values) for the cell-type to illustrate how both can help determine TF-TF behaviors involved in gene regulation of TG. The final NetREm model predicted the main actions of TFs 1, 2, 4, and 5 in gene regulation of target gene TG. TF_3 is weakly involved. It is likely that from the B matrix of TF-TF interactions we may predict: Cooperative behavior (+): TF1 and TF2 (and weakly with TF3); TF4 and TF5. That is, cooperative behavior ($B > 0$) is among co-activators and co-repressors since they have the same goals for TG: co-activators want to increase transcription and subsequent expression of TG; co-repressors want to prevent or reduce transcription of TG to messenger RNA (mRNA). Antagonistic behavior ($B < 0$): TF_1 and TF_4; TF_1 and TF_5; TF_2 and TF_4; TF_2 and TF_5; TF_3 and TF_4; TF_3 and TF_5. NetREm's model predicts that there is antagonism between an activator and repressor because both have opposing behavior to regulate TG.

Activator modules and repressor modules and these form transcriptional regulatory modules (TRMs) together. While the figure shows that they may bind to different Regulatory Elements (REs), they can all bind to the same ones as well. Nonetheless, NetREm can be applied to many applications where an inherent network structure exists among the predictors (e.g. TFs) to influence the response variable y (the expression levels of our TG that may be regulated by some of the N candidate TFs). The figure on the right shows how cell-type TF-TF coordination networks $\bar{B}$ may differ. Different cell-type TFs may interact with each other differently across cell-types to regulate the TG.



Section §B.1.4: SERGIO simulator for Human Embryonic Stem Cells (hESCs)

We randomly select 70% of data for training so M = # of train cells. We randomly select 1,250 TGs and corresponding TFs from weighted and signed (+: activates; -: represses) ground truth GRN atlas from TF induction analysis on expression and ChIP-seq studies (Sharov et al. 2022). This results in $\mathcal{N} = 207$ TFs (192 master regulators) and 5,050 signed GRN ground truth links we input to the SERGIO (Single-cell ExpRession of Genes In silicO) tool (Dibaeinia and Sinha 2020; Dibaeinia 2024) to simulate realistic single-cell data for 1,000 cells (M =

700) and 1,442 genes. We vary noise (30%, 60%, 90%), retrieving 3 different synthetic expression datasets. We repeat this for 100 cells (M = 70).

§B.1.4.1: Motivation for using the SERGIO tool

It is challenging to find signed ground truth gene regulatory networks (GRNs) for evaluating c^*-TF-TG regulatory networks. Other real-world applications with gold standards may lack sign information and only provide TF-TG regulatory links, or the gold standards may be measured in different experiments with varying assumptions, making it unreliable for comparing the coefficients' signs. Therefore, we want to evaluate not only our TF-TG regulatory links but also the inferred signs. Additionally, we aim to test different sample sizes M (number of cells) and noise ratios to assess our performance. SERGIO is a single-cell simulator tool that generates realistic gene expression datasets by incorporating TF-TG regulatory interactions, such as the stochastic nature of transcription and regulation of TFs by multiple TFs. Data simulated by SERGIO is shown to be comparable to experimental data generated by Illumina HiSeq2000, Drop-seq, Illumina 10X chromium, Smart-seq.

We use SERGIO to evaluate and benchmark NetREm's performance on gold standard data for hESCs. Our goal is to measure and benchmark our TF-TG regulatory networks holistically, including the predicted sign (c^*: + or -) because these signs are used to assign TF regulatory roles. Specifically, $c^* > 0$ indicates an activator and $c^* < 0$ indicates a repressor. NetREm also uses these signs to infer the nature of potential coordination among TFs for regulating the TG: $B < 0$ indicates an antagonistic relationship, $B > 0$ indicates a cooperative relationship.
Thus, we use SERGIO to obtain single-cell gene expression data for our hESC ground truth GRN. This allows us to run NetREm and other benchmark regression models (BRMs: LassoCV, RidgeCV, ElasticNetCV, Linear Regression) to compare their accuracy, sensitivity, and specificity. SERGIO generates datasets that are statistically comparable to experimental data, modeling the stochastic nature of gene regulation and the role of many TFs in regulating TGs. This tool enables us to reverse engineer by building and inputting a ground truth GRN, and then generating a realistic gene expression dataset based on that. SERGIO is used by other studies to help evaluate and benchmark the performance of various GRNs. By using SERGIO, we ensure that the expression dataset is constructed to align realistically with the ground truth GRN, allowing us to control for other sources of variability and focus solely on the relative performance of different tools for GRN inference.

§B.1.4.2: Signed ground truth gene regulatory network (GRN) for hESCs:

We obtain an atlas of regulated TGs of TFs in hESCs (Sharov et al. 2022), and analyze results provided in Additional File 5 of the study (Table of all regulated targets of TFs). This table focuses on 94,395 TF-TG regulatory links in hESCs learned after combining 2 data sources. 1: TF binding data (e.g. ChIP-seq data). 2: data on changes in the expression levels of genes (i.e. increase or decrease and by how much) after the given TF is induced so its expression level and presence in the cell is boosted (i.e. how does the TF expression impact expression levels of other genes). To control the False Positive (FP) rate (predicted TF-TG regulatory links that may be false), the study utilizes the Expected Proportion of False Positives (EPFP) to guarantee the FP rate is below a specified threshold. The table provides the direction of TG expression change after TF induction and the corresponding Log10-ratio of TG expression change: $\Delta = \log_{10}\left(\frac{TG \text{ expression before TF induction}}{TG \text{ expression after TF induction}}\right)$ where $\text{direction} = \begin{cases} down \text{ if } \Delta < 0 \\ up \text{ if } \Delta > 0 \end{cases}$. For simplicity, the table focuses on the dominant TG expression change direction, where it predicts that when direction is down, the TF represses the TG (i.e. $\Delta < 0$ → TF decreases TG expression→ coefficient can be assumed negative) and when direction is up, the TF activates the TG (i.e. $\Delta > 0$ → TF increases TG expression → coefficient can be assumed positive) (Sharov et al. 2022).

There are other columns of data such as the Direct-EPFP, which is N/A if there is no ChIP-seq data. We retain highly-confident TF-TG regulatory links that have ChIP-seq support. Thus, we remove rows with no ChIP-seq data, resulting in 57,272 rows of TF-TG regulatory links comprising 14,079 genes: 14,044 TGs, 266 TFs (of which, 35 TFs are master regulators and hence are not TGs). We did not have any underlying gene expression data for steady-state human ESCs, but instead use this table as a ground truth GRN, where edges are directed from TF and TG and are weighted and signed. Here, weight is given by the $\log_{10}$(ratio) of TG expression change after TF induction. Out of the 14,044 TGs, we randomly selected 1,250 unique TGs. Then, we select their corresponding TFs based on the ground truth GRN. Based on our randomly selected TGs, we ultimately select 1,442 genes in total, comprising 207 TFs. Since we subset 1,250 TGs and are only adding their direct TFs, we miss other regulatory TF-TG interactions and hence 192 of those 207 TFs are considered master regulators in our subset GRN (although some of them may be TGs for other TFs not in the list of 1,442 genes).

§B.1.4.3: Generating gene expression data based on input ground truth GRN:

We utilize SERGIO to simulate single-cell gene expression data guided by our underlying ground truth GRN. A key advantage of SERGIO is that the user can alter and tune different parameters to customize the resulting gene

expression data set. We can adjust various parameters for SERGIO, such as the number of genes (we set number_genes = 1442), number of cells (total number of cells per cell-type to be simulated; we set number_sc = 1000), noise type (The type stochastic noise for genes; we set: dpd for Dual Production Decay), sampling state (length of simulations in stationary region; we set: 15), number of cell types (total number of distinct cell types to be simulated; we set number_bins = 1), decays (decay parameter for genes in steady-state simulations; we set: 0.2). For our GRN structure, we input the log10 values for our maximum interaction strength where negative values are for repressive TF-TG interactions and positive values are for activating interactions. We assume that all 192 master regulator TFs have the same basal production rate in the cell-type, so we give a value of 1 for each. We use shared_coop_state = 2, where this value (used for all hill coefficients in simulations) is recommended to be between 1 and 3. We retain all other defaults but vary noise_params to generate 3 different gene expression datasets. That is, we adjust the noise parameter (noise_params) to be 0.3, 0.6, and 0.9, where noise_params controls the noise amplitude parameter q for the genes in our steady-state simulations. SERGIO finds that the accuracy of GRN network inference is impacted greatly by technical noise, and we want to evaluate NetREm's performance and robustness to these real-world problems. This first setting outputs 3 gene expression datasets for 1,442 genes and 1,000 cells (single-cell samples). Since we hold out 30% of gene expression data for testing and use the remaining 70% for training, we have $M = 700$ cells. In this case, $N = 207 < M = 700$.

We also explore scenarios where $M < N$, to benchmark NetREm's performance in a scenario where there are more predictors and fewer samples. Hence, we adjust SERGIO so that we also create gene expression datasets for 100 samples (training: 70 cell samples, testing: 30 cell samples). Thus, we retain the previous settings but set number_sc = 100. We still vary the noise parameters (30%, 60%, 90%). Hence, SERGIO returns 6 gene expression datasets for human embryonic stem cells:

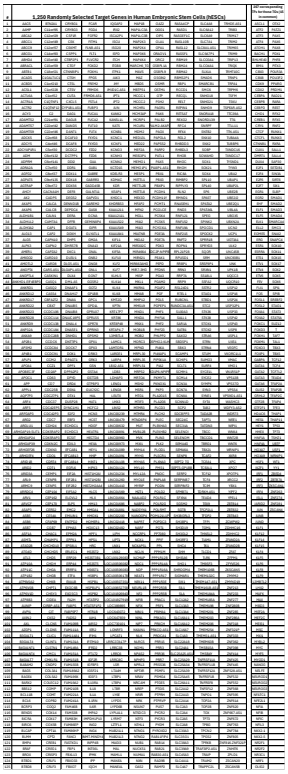
- $M > N$: 1,000 cells: $M = 700$, N = 207 TFs (206 if TG is also a TF). Noise parameter: 30%
- $M > N$: 1,000 cells: $M = 700$, N = 207 TFs (206 if TG is also a TF). Noise parameter: 60%
- $M > N$: 1,000 cells: $M = 700$, N = 207 TFs (206 if TG is also a TF). Noise parameter: 90%
- $M < N$: 100 cells: $M = 70$, N = 207 TFs (206 if TG is also a TF). Noise parameter: 30%
- $M < N$: 100 cells: $M = 70$, N = 207 TFs (206 if TG is also a TF). Noise parameter: 60%
- $M < N$: 100 cells: $M = 70$, N = 207 TFs (206 if TG is also a TF). Noise parameter: 90%

Theoretically, 258,735 TF-TG regulatory links are possible and only 5,050 of these links will be true. Thus, we adjust the noise parameter (we set noise_params = 0.3, 0.6, 0.9) and the # of cells to have these 6 datasets. We will

be able to also evaluate the coefficients c^* for the TF predictions (where $c^* > 0$ is for activators and $c^* < 0$ is for repressors) since our input GRN is weighted and signed (and is used to generate respective gene expression data).

§B.1.4.4: Corresponding human input protein-protein interaction (PPI) network (PPIN):

To estimate the number of true negatives (TN) using the number of TFs and TGs, we make some assumptions about the interactions within this ground truth GRN. The total possible interactions can be calculated if we assume that every TF can potentially regulate every TG. However, the presence of overlaps between the TFs and TGs introduces a slight complexity, which we will address. So, out of 1,250 TGs and N = 207 TFs (of which 15 are also TGs), we find there are 258,735 total possible TF-TG links (where TF ≠ TG). We benchmark NetREm's performance for constructing TF-TG regulatory links for the 1,250 TGs in hESCs across each of the 6 SERGIO-simulated datasets. We input the comprehensive human PPI network of 21,321 edges comprising 10,777 known links and 10,544 artificial links (default edge weight of 0.01) for these 207 TFs. For each of the 15 TFs that are also TGs, we remove the respective TF and its associated edges from the input PPI network and node that $N = 206$ for those TFs; otherwise, $N = N^* = 207$ for the remaining 1,235 TGs, since we do not incorporate other GRN knowledge to select candidate TFs.



Section §B.1.5: Gene Expression Data for 7 Main Applications (Apps)

In applications 1, 6, 7, we randomly select 70% of data for training so M = # of train cells. We normalize and scale data with Seurat (Satija et. al 2015).

§B.1.5.1: human Embryonic Stem Cells (hESCs)

We obtain a weighted and signed (+: activates; -: represses) atlas of ground truth TF-TG regulatory information in hESCs, originating from TF induction analysis on gene expression and from ChIP-seq studies (Sharov et al. 2022). We randomly select 1,250 TGs and their TFs from the ground truth GRN, which results in $N^* = 207$ TFs (15 that are also TGs, 192 that are master regulators) and 5,050 signed TF-TG links that we input to the state-of-the-SERGIO (Dibaeinia and Sinha 2020; Dibaeinia 2024) tool that then simulates realistic single-cell gene expression data for 1,000 cells (training: M = 700, testing: 300) and 1,442 genes based on this underlying, true GRN subset. We vary SERGIO's noise parameter (30%, 60%, 90%) and ultimately retrieve 3 different synthetic single-cell gene expression datasets for hESCs. We repeat this for 100 cells (M = 70).

§B.1.5.2: human Hematopoietic Stem Cells (HSCs)

We utilize Seurat to normalize and scale single-cell gene expression data for 12,558 genes in HSCs (Buenrostro et al. 2018). We remove genes not expressed in ≥ 3 cells in the total data, resulting in 12,223 genes for 2,268 cells in HSCs. We focus on running NetREm for TGs that with prior ground truth GRN validation data (Zhang et al. 2023). Hence, the final data has 10,588 genes for $M = 2{,}268$ cells in HSCs. We use all data for training and for testing. We also note there are $\mathcal{N} = 178$ TFs found in both the ground truth GRN and the gene expression data, which are used as the candidate TFs for each TG. For each TG, N = 177 (if the TG is also a TF) or 178 TFs (if the TG is not a TF).

§B.1.5.3: mouse Embryonic Stem Cells (mESCs)

We use data (GSE108222) from a study (Tran et. al 2019) that reprograms mouse embryonic fibroblasts (MEFs) to an embryonic-like induced pluripotent stem cell (iPSC) state (i.e. "turning back the clock" on the cellular identity) by reversing cellular differentiation end epigenetic markers of aging. Since they pass all pluripotency tests, iPSCs have a pluripotent state functionally equivalent to that of mESCs. We follow (Zhang et al. 2023) and concatenate the expression of 2 mice replicates at each time point. We use Seurat (Satija et. al 2015) to normalize and scale the gene expression data. Then, we filter the 24,421 genes to include only TGs and TFs found in ground truth data. Our final data has 19,225 genes for M = 1,080 cells in mESCs. We also note there are $\mathcal{N}$ = 195 TFs found in both the ground truth GRN and the gene expression data, which are used as the candidate TFs for each TG. Final data has 19,225 genes for M = 1,080 cells. N = 194-195 candidate TFs for each TG.

§B.1.5.4: mouse Dendritic Cells (mDCs)

We use normalized data (McCalla et al. 2023) originating from study GSE48968 (Shalek et. al 2014) on single-cell RNA-seq libraries for over 1,700 primary DCs derived from bone marrow in mice. Ultimately, we find $\mathcal{N}$ = 93 TFs in both ground truth GRN and expression data, which are candidate TFs for each of the 9,087 TGs. This expression data has M = 1,211 cells. N = 92-93 for 9,087 TGs for M = 1,211 cells.

§B.1.5.5: human Peripheral Blood Mononuclear cells (PBMCs)

Using publicly available 10X Genomics data on 2,700 PBMCs from a healthy donor, we follow (Satija et. al. 2024) to retrieve 9 immune cell subpopulations with 13,714 TGs each. We use 1,639 human TFs (Lambert et al. 2018). We find 1,029 TFs in this list are also in the gene expression data.

Cell-type	N	M	Maximum Condition # across TGs
PBMC Naive_CD4_T-C0	1,029	709	13.8
PBMC CD14_Mono_C1	1,029	480	12.7
PBMC Memory_CD4_T-C2	1,029	429	21.9
PBMC B_C3	1,029	342	22.8
PBMC CD8_T_C4	1,029	316	25.6
PBMC FCGR3A_Mono_C5	1,029	162	35.6
PBMC Natural_Killer_C6	1,029	154	59.9
PBMC Dendritic_Cell_C7	1,029	32	425
PBMC Platelet_C8	1,029	14	451

§B.1.5.6: Human myelinating (mSCs) and non-myelinating (nmSCs) Schwann cells (SCs)

We utilize Seurat to normalize and scale gene expression datasets for 13,886 genes in mSCs and nmSCs in the Dorsal Root Ganglion (DRG) L4,5 regions for 5 human donors (Avraham et al. 2022). We include only protein-coding genes and remove genes not expressed in at least 3 cells in the total data (comprised of other cell types). Our final data has 17,049 genes for 319 cells (training: M = 223, testing: 96) in mSCs and 2,468 cells (training: M = 1,727, testing: 741) in nmSCs.

§B.1.5.7: Alzheimer's disease (AD) vs. Controls in 8 Central Nervous System (CNS) Cell-Types

Astrocytes (Astro), Oligodendrocytes (Oligo), OPCs (Oligodendrocyte Precursor or Progenitor Cells), Microglia (Mic), Inhibitory GABA-ergic Neurons (InNs), Excitatory Glutamatergic Neurons (ExNs), Pericytes (Peri), Endothelial Blood Brain Barrier (BBB) cells (Endo. BBB). We use the processed data for 24 individuals with AD pathology and 24 healthy controls provided by Gupta et. al (Gupta et al. 2022) for all 8 neuronal/glial cell types. This data is from the original AD study (Mathys et. al 2019) of 80,660 droplet-based single-nucleus transcriptomes (i.e. the cells) for 17,926 genes from the prefrontal cortex for 8 neuronal/glial cell types. In AD, we have the following number of cells: Astro (1,830, $M = 1{,}281$ train, 549 test), Oligo (9,035, $M = 6{,}324$ train, 2,711 test),

OPCs (1,290, $M = 903$ train, 387 test), Mic (955, $M = 668$ train, 287 test), InNs (4,371, $M = 3{,}060$ train, 1,311 test), ExNs (17,878, $M = 12{,}515$ train, 5,363 test), Pericytes (76, $M = 64$ train, 27 test), Endo. BBB (59, $M = 41$ train, 18 test). These are the respective counts in controls: Astro (1,562, $M = 1{,}093$ train, 469 test), Oligo (9,200, $M = 6{,}440$ train, 2,760 test), OPCs (1,337, $M = 936$ train, 401 test), Mic (965, $M = 676$ train, 289 test), InNs (4,825, $M = 3{,}378$ train, 1,447 test), ExNs (17,098, $M = 11{,}969$ train, 5,129 test), Pericytes (91, $M = 53$ train, 23 test), Endo. BBB (62, $M = 19$ train, 43 test).

Section §B.1.6: Gene Expression Data for Validation

§B.1.6.1: Sciatic Nerve (SN) Atlas Data for Schwann Cells (SCs) in Mice

We utilize raw RNA-seq data for mSCs and nmSCs in the SN of 4 mice from GSE137947 (Gerber et. al 2021), which comprises the SN Atlas (SNAT). We use g:Profiler(Reimand et al. 2007) orthologous functionality to help map the mouse (mus musculus) gene IDs to 16,954 human (homo sapiens) gene names. The gene expression data has data for 17,048 genes in myelinating (mSCs) and non-myelinating (nmSCs) SCs. This dataset helps prune down our list of TFs in mSCs and nmSCS, when we define our prior gene regulatory knowledge.

§B.1.6.2: Genotype-Tissue Expression (GTEx) Data for SCs

We utilize this as a validation dataset to evaluate NetREm in human Schwann cells. This GTEx data from (Eraslan et. al 2022) for SCs pooled from 5 tissues (esophagus mucosa/muscularis, heart, prostate, skeletal muscle). After applying Seurat and MAGIC normalization, this dataset contains 14,144 genes across 1,430 SCs.

§B.1.6.3: Gene Expression data for 4 Neuronal/Glial Cells

This helps evaluate NetREm for 4 neuronal/glial cells. (Lake et al. 2018) provides data for 4 Central Nervous System (CNS) cell types: Excitatory neurons (ExNs), Inhibitory neurons (InNs), Microglia (Mic), Oligodendrocytes (Oligo). This undergoes preprocessing according to methods outlined in (Jin et al. 2021): removing genes expressed in <100 cells, normalization and scaling (Satija et. al. 2024), imputation using MAGIC algorithm. Final data comprises 15,692 genes across 13,709 ExNs, 6,045 InNs, 317 Microglia, 2,657 Oligodendrocyte cells.

Section §B.1.7: Prior Knowledge on Gene Regulation

Please note that we utilize R Statistical Software and Bioconductor packages to derive several databases that are used in this analysis.

Prior gene regulatory network (GRN) knowledge on gene regulation (overview)

In human applications 6 (Schwann cells (SCs)) and 7 (Alzheimer's disease (AD)), we build and use prior reference GRN knowledge that subsets N biologically-promising candidate TFs for each TG in the cell-type (N can differ across TGs). Recent data enables the construction of prior GRNs for various cell-types in the human body (e.g. (Zhang et al. 2021) uncovers regions of open chromatin in 222 human cell-types). To construct prior candidate TF-Regulatory Element (RE)-TG regulatory links (prior reference GRNs), we use a multi-step approach based on available data (e.g. TF binding predictions, chromatin accessibility). Incorporating multi-omics data can lead to more biologically meaningful and potentially truer cell-type TF-TG regulatory networks (Badia-i-Mompel et al. 2023). Such sources help determine N candidate TFs for each TG, whose expression levels define our input X when running NetREm for that TG. We provide step-by-step details and explanations for these 4 steps as a potential guide:

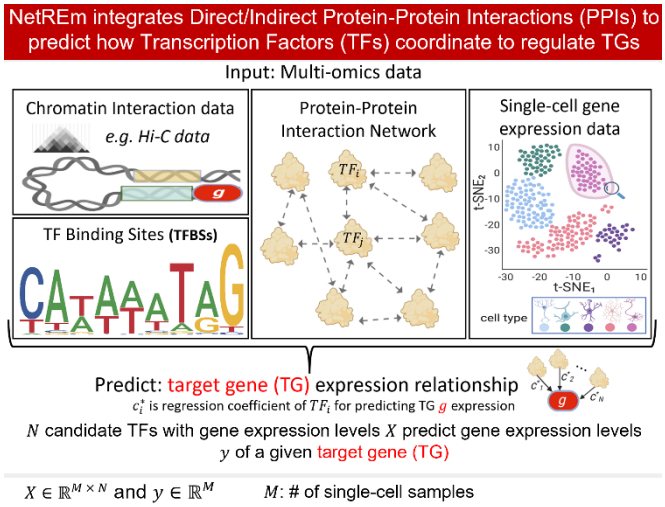
1. Map cell-type REs to TGs they may be associated with: we identify raw scATAC-seq peaks (regions of open, accessible DNA chromatin) in adult cell types (Zhang et al. 2021) relevant to our study. We obtain various annotated databases of RE to TG links and map these non-coding scATAC-seq peaks to TGs by overlapping them with RE-TG links. This results in peak-to-TG predicted links. By focusing on cell-type chromatin accessibility, we consider TGs open to transcription in that cell type and regulatory regions where TFs may potentially bind to regulate them, which is crucial for predicting TF-TG links (Badia-i-Mompel et al. 2023).

2. Motif-based GRN identifies potential TFs binding to REs to regulate their TGs: We predict TF binding using motif matcher algorithm tools (motifmatchr (Schep and University 2023) and monaLisa (Machlab et al. 2022)) along with Position Weight Matrix (PWM) databases to scan the RE regions. Generally, a given motif binds to a single TF motif at a time, with exceptions for dimeric transcription factors (e.g., AP-1, JUN/FOS). We predict TFs in this motif-based GRN that recognize and bind to their sequence-specific motifs on TFBSs of various REs (e.g., promoters, enhancers, eQTL-based regions), thereby regulating the expression of their respective TGs. This results in a list of motif-GRN TFs for each TG, revealing paralogs and TFs recognizing the same motif and competing for TFBSs (Berenson et al. 2023).

3. Pruning Motif-based GRN TFs: We chop motif-GRN TFs with relatively lower motif-matching scores (based on output from motif matcher algorithm tools) and remove TFs with relatively low expression in training expression data (and/or external data for same cell-type).

4. TF Augmentation Step (Consider Potential Colocalization among TFs): Finally, since solely relying on PWMs to model motifs and predict candidate TFs for TGs has drawbacks, we implement an augmentation step. These drawbacks include assuming independence of positions within a binding motif (Siddharthan 2010), low specificity of many TF binding motifs (Badia-i-Mompel et al. 2023), the fact that many TFs share similar binding motifs (Ali et al. 2022). These issues lead to most ChIP-seq peaks for TFs missing relevant PWM hits for corresponding TFs (Tsukanov et al. 2022). TF binding to a TG does not always imply regulation, as TFs often cooperate with other molecules to regulate TGs (Siddharthan 2010). TF regulatory mechanisms involve other factors besides accessibility of TF binding motifs, such as pioneer TFs (Kamimoto et al. 2023). Therefore, we add TFs based on criteria like potential colocalization with our pruned TFs, forming complexes, or exhibiting strong molecular functional similarity. We check these TF pairs have known PPI links (and are not artificial), resulting in N customized, candidate TFs for a given TG.

Goal: Deriving Prior Gene Regulatory Network (GRN) for each Target Gene (TG)



Please note the approaches mentioned below are examples of how to derive a prior reference GRN for the input step for NetREm. Performing this step is optional but this initial feature selection may lead to more biologically relevant results. In addition, these steps may be adapted and adjusted based on personal preferences or application needs. In short, there are many ways and various existing tools that can be utilized to find a list of N candidate Transcription Factor (TF) # for each target gene (TG). Each TG may have a different final number of candidate TFs, so N may differ across TGs.

§B.1.7.0 Organizing Input Data

§B.1.7.0A Raw single-cell ATAC-seq regions

We obtain raw single-cell ATAC-seq peaks for 222 cell types (Zhang et al. 2021), corresponding to regions of open and accessible DNA chromatin (euchromatin) in non-coding DNA. We focus on relevant adult cell-types.

§B.1.7.0B Determining cell-type-specific single-cell ATAC-seq regions

The following details how we determine cell-type-specific single-cell ATAC-seq peaks (regions of open and accessible DNA chromatin) based on the raw scATAC-seq data from **§B.1.7.0A**. Below, we explain our process for SCs, which we applied for the other adult cell-types for our AD application. We note that this process can be applied to other cell-types of interest, based on the data that is available.

Schwann cells (SCs): We focus on the raw SC ATAC-seq file, which has 88,175 rows corresponding to 73,114 unique peaks (chrom #: start position - stop position). Our goal was to find ATAC-seq peaks that are mainly specific to Schwann cells and/or potentially limited to up to these 6 others closely-related cell-types: fetal Schwann cells, fetal enteric glia, fetal Oligo Progenitor 2, adult melanocytes, adult oligodendrocytes, adult oligodendrocyte progenitor cells (OPCs). To determine any and all overlaps between Schwann cell ATAC-seq peaks and the peaks of each of the other 221 cell types, we employed the GRanges (Lawrence et. al 2013) package in R. This allows for efficient identification of the shared genomic regions between the SC peaks and the peaks of the respective cell types. Thus, for each of the 221 cell types, we overlapped their respective raw ATAC-seq peaks with those of the general SC raw ATAC-seq peaks. For a given cell type, we use the 2-sided Fisher's exact test (via fisher.exact in R) to assess the statistical significance of each overlap identified between a SC ATAC-seq peak and that cell type's ATAC-seq peak. This was preferable to defining a strict overlap cutoff number of bases since raw ATAC-seq peaks had differing widths. By comparing the observed overlap with the expected overlap under independence (if the overlap is not statistically significant), Fisher's exact test provided a p-value indicating the likelihood of obtaining such an overlap by chance (i.e. is observed overlap higher or lower than expected). To account for multiple testing, a Benjamini-Hochberg multiple test correction was applied. This adjusts the p-values for each overlap to control the false discovery rate. Only overlaps with an adjusted p-value (p_adj) < 0.05 are considered statistically significant between SCs and the respective cell type. Those significant peaks were then annotated as overlaps between SCs and that cell type and the non-significant overlaps (p_adj > 5%) were discarded. We repeat this process for all 221 cell types. Ultimately, SC-type-specific scATAC-seq peaks are only found in raw SCs and do not have any statistically significant overlaps with any of the ATAC-seq peaks for the other 221 cell types. We found 1,736 such peaks. Nonetheless, there are strong shared biological signals between Schwann cells and those 6 cell types mentioned above. Hence, we expanded our scope and loosened our criteria (to uncover more Schwann cell ATAC-seq peaks) to consider Schwann cell-type-specific scATAC-seq peaks as those that may significantly overlap with any of those 6

cell-types only and do not overlap with any of the other 215 cell types. That added 1,052 peaks. Finally, we uncover 2,788 such Schwann cell ATAC-seq peaks.

Alzheimer's disease (AD): We perform this analysis for the 8 other cell-types for Alzheimer's disease (AD). For Endothelial Blood Brain Barrier (BBB) cells, we consider similar cell-types as: Endothelial exocrine, endothelial general 1-3, endothelial myocardial, fetal endothelial general 1 – 3, fetal endothelial hepatic 1 and 2, fetal endothelial placental. For Endothelial cells overall, we consider similar cell-types as Endothelial exocrine, endothelial BBB, endothelial myocardial, fetal endothelial general 1 – 3, fetal endothelial hepatic 1 and 2, fetal endothelial placental. For Inhibitory Neurons, we consider similar cell-types as fetal Inhibitory neurons 1 to 5. For excitatory neurons, we consider similar cell-types as fetal excitatory neurons 1 to 12. For OPCs, we consider fetal oligo progenitor 2. For Oligodendrocytes, we consider Fetal Oligo progenitor 2, Oligo Precursor, fetal SC general, and SC general. For Astrocytes, we consider fetal Astrocytes 1 to 5. For Microglia, we consider fetal Macrophage general 1 to 4, fetal macrophage hepatic 1 to 3, fetal Macrophage placental, Macrophage general or alveolar, Macrophage general. For Pericytes, we consider: Pericyte muscularis and cardiac pericyte 1 to 4.

§B.7.0C Obtaining peak-to-TG predicted links

To construct prior candidate TF-Regulatory Element- (RE)-TG regulatory links (prior reference GRNs), we use a multi-step approach based on the available multi-omics data. We start by identifying raw scATAC-seq peaks from **§B.1.7.0A**. We map these peaks to TGs by overlapping them with annotated RE to TG links. This process entails using various approaches to obtain these peak-to-TG predicted links, which we explain below as a potential guide.

1. We gather a database of cis-Candidate REs (cCREs) (Zhang et al. 2021) that are already mapped to TGs based on genomic distance limits(Badia-i-Mompel et al. 2023) and previously filtered based on their Activity by Contact (ABC) scores. We use all interacting cCREs for that cell-type (e.g. promoter-promoter, promoter proximal – promoter, distal (enhancer) – promoter) that are present in adult tissues. These various types of chromatin interactions may illuminate different aspects of coordination among TFs; for instance, studies observe that since promoter-promoter interactions may cluster several TGs, they can provide unparalleled information on TF coordination, especially among TGs that may be context-specific (e.g. tissue- or cell-type-specific) or involved in cellular housekeeping processes (Li et al. 2012). We thus learn the following potential regulatory elements for each TG: gene promoters, cCREs (promoter, promoter-proximal, and/or distal).

2. We use TxDb.Hsapiens.UCSC.hg38.knownGene (Team BC, Maintainer BP, et. al 2019; Zhou et. al 2018 2018) to define default promoter-only regions and then overlap those with scATAC-seq peaks.
3. For Oligo, Mic, and Neurons (ExNs and/or InNs), we also utilize Plaq-seq data(Nott et al. 2019) on promoter-enhancer interactomes and enhancer regions in hg19. We run scGRNom(Jin et al. 2021) Step 1 (scGRNom_interaction) on this epigenomic data to predict all possible interactions between enhancers and promoters and annotate them with the associated TGs. We use rtracklayer (Lawrence et al. 2009) and the hg19tohg38.over.chain file from UCSC (Karolchik et al. 2009) to perform liftover(2021) to map the associated enhancer and promoter regions to hg38 human genomic coordinates. We use these annotated interaction enhancer and promoter regions as Plaq-seq-based regulatory regions for the respective cell-types.

We overlap our raw scATAC-seq peaks with these annotated regulatory elements to TG links to associate the scATAC-seq peaks with their potential TGs.

4. Then, we look at the remaining scATAC-seq peaks that did not map to any regulatory elements and attempt to characterize them via expression quantitative trait (eQTL) data that links single nucleotide polymorphisms (SNPs) to TGs whose expression levels they impact. We use GTEx v8 Tibial Nerve eQTL data (the left peripheral tibial nerve from the gastrocnemius region) for SCs (THE GTEX CONSORTIUM 2020) and eQTL data from Byrois et al. for brain cells for our AD application (Bryois et al. 2022).

§B.1.7.0D Position Weighted Matrices (PWMs) of TF Binding Motifs

Different TF binding motif databases have different coverages of TFs(Badia-i-Mompel et al. 2023) and diverse annotated binding sites(Skok Gibbs et al. 2022). We obtain PWMs for humans from R Bioconductor (Gentleman et al. 2004) databases. Each motif is specific to 1 TF and a TF may have many motifs. When running motifmatchr (Schep et. al 2023) to search for motif matches for TFs, we use JASPAR2022 (Castro-Mondragon et al. 2022) for humans (species = 9606) and convert Position Frequency Matrices (PFMs) to PWMs using TFBSTools (toPWM() function, type = “prob”(Tan and Lenhard 2016)). We have 949 PWMs corresponding to 680 TFs. To run monaLisa (Machlab et al. 2022) for motif analysis and TF binding prediction, we use a comprehensive database of 8,903 PWMs (last update: March 4, 2022). We use MotifDb (Shannon and Richards 2023) for “hsapiens” to obtain 6,086 PWMs and motifbreakR’s motifbreakR_motif data (Coetzee et al. 2015) to add 2,817 PWMs. We remove redundant motifs (found in both sources) for TFs when running monaLisa for our analysis.

Since DNA strands are antiparallel, they run in opposite directions where one strand runs 5’ to 3’ while the complementary strand runs 3’ to 5’. A palindrome in DNA refers to a sequence of base pairs that reads the same in the 5' to 3' direction on one strand as it does in the 5' to 3' direction on the complementary strand when both strands are aligned. In regulatory DNA, palindromes are inverted sequence repeats (read the same from 5’ to 3’ on complementary DNA strands). These cis-regulatory palindromic motifs can be bound, cooperatively, by homodimers (same TF: $TF_i :: TF_i$) or by heterodimers (different TFs: $TF_i :: TF_j$) to mediate cell type-specific combinatorial gene regulation. That is, a sequence-specific TF can bind to the palindromic motif on opposing DNA strands as a homodimer, where 1 TF binds with its DNA-binding domain to the left and another copy of the TF binds to the right half-site of the motif(Datta and Rister 2022). Thus, if TFs bind as homodimers, then the motif in PWMs will reflect this behavior. For instance, JUN-family TFs have a palindromic AP-1 motif (TGAG/CTCA), which enables them to form homodimers with other JUN TFs or form heterodimers with FOS family TFs. Our current approach looks at coordination among different TFs. To incorporate homodimers in our analysis, we need robust prior knowledge on homodimeric TF-TF interactions and need to adjust NetREm to handle self-interactions.

§B.1.7.1 Constructing Motif-based Gene Regulatory Networks (GRNs)

Essentially, we predict TF Binding Sites (TFBSs) for TFs. We utilize the peak-to-TG predicted links for each TG from **§B.1.7.0C**. Then, we predict candidate TFs that may bind to these regulatory regions for the TG based on their sequence-specific TFBS motifs. To this end, we utilize 2 different tools to construct these motif-based GRNs: monaLisa and scGRNom.

§B.7.1A Utilizing monaLisa

To predict strong TF binding to the regulatory regions for TGs, we run monaLisa with a very stringent minimum motif matching score of 8 with the “matchPWM” method. We use our comprehensive PWM of 8,903 PWMs. For TF binding to regulatory regions, we use method = “matchPWM” and min.score of 8 to identify more TFs binding to the regulatory element (RE) for each of the cell-type-specific REs that we determined. We use our more comprehensive input Position Weighted Matrix (PWM). The monaLisa tool returns TFs predicted to strongly bind to each given regulatory region along with their scores and predicted positions in the regulatory region. This is our monaLisa-motif-based GRN that links TFs to REs (and specific positions within them) to the target genes (TGs). For instance, these regulatory elements have chromosome #: start position – end position.

§B.7.1B Adapting scGRNom

We use the JASPAR2022 database (Castro-Mondragon et al. 2022) and hg38 human reference genome. We adapt the scGRNom(Jin et al. 2021) pipeline for building cell-type gene regulatory networks (GRNs) such that we run scGRNom_getTF (Step 2) on all the regulatory regions for the TG, including promoters, eQTL-based regions, proximal promoters, distal (enhancers). We use all the defaults mentioned by scGRNom tool to predict TFs regulating the TG based on scGRNom. We output a reference scGRNom-motif-based GRN that links TFs to regulatory elements to the TGs.

§B.7.1C Combining the motif-based GRNs

Thus, for a given TG, we run monaLisa in **§B.1.7.1A**, to identify a pool of TFs that are predicted to regulate this TG based on our monaLisa binding prediction (using the comprehensive PWM) run on all regulatory elements we have mapped to this TG. We also find the list of TFs in **§B.1.7.1B** predicted by scGRNom Step 2 (on JASPAR2022 database (Castro-Mondragon et al. 2022)) to regulate this TG based on the regulatory elements that have been mapped to this TG. Then, we look at all the TFs that are identified in **§B.1.7.1A** and/or **§B.1.7.1B**, and we add in any TFs that they may be found to form a protein complex with (i.e. score in input Protein-Protein Interaction (PPI) network of 1). For instance, if FOS is identified as one of the potential TFs for the TG (based on **§B.1.7.1A** and/or **§B.1.7.1B**), then we ensure that JUN is also added to the list of TFs (if it is not already in the list of TFs), since FOS and JUN are known to form a complex FOS::JUN. Thus, **§B.1.7.1C** results in a pooled list of TFs from **§B.1.7.1A** and/or **§B.1.7.1B** and the associated TFs that they form known TF complexes with. Going forward, please note we explain this process for a given TG in the cell-type (or tissue-type) for the particular context (e.g. control, disease). This process will be applied for each TG in the cell-type.

§B.1.7.2 Pruning TFs down based on expression levels

This is the reduction stage, where we will prune down the list of TFs that we retrieve in **§B.7.1C**, for the TG, based on their relative expression gene expression levels. That is, key drivers of cell-type-specific gene regulation ought to have relatively meaningful expression levels in the cell **(§B.7.2A-§B.7.2B)**.

§B.1.7.2A Removing TFs based on global criteria for all genes in the dataset

We will remove TFs from **§B.7.1C** that have a relatively lower level of gene expression in the cell-type (or tissue-type) compared with other genes. Otherwise, we will use our training gene expression dataset for this part (not our testing data). If we have another gene expression dataset available for a similar context (e.g. cell-type or tissue-

type), we can use that external dataset as well. That is, we will look at that (training) gene expression data and prune out TFs based on their expression levels relative to the expression levels of all other genes in the dataset. Below, we detail how we do this for SCs (where we do have external gene expression data we can use) and for AD versus controls. For the AD application, since we only have external data (Lake et al. 2018) for 4 out of the 8 cell-types, we opt not to use this external dataset, to help keep the process consistent for the 8 cell-types.

Schwann Cells

We also utilize raw RNA-seq data for mSCs and nmSCs in the Sciatic Nerves (SN) of 4 mice from GSE137947 (Gerber et. al 2021), which comprises the SN Atlas (SNAT). We use g:Profiler (Reimand et al. 2007) orthologous functionality to help map the mouse (mus musculus) gene IDs to 16,954 human (homo sapiens) gene names.

- **Step 1:** We identify the 40th percentile for all 17,048 genes in the training gene expression dataset in the Dorsal Root Ganglion (DRG).
  - In human mSCs, this 40% threshold is: 2.9e-2. Here, 6,819 genes are below this threshold.
  - In human nmSCs, this 40% threshold is: 2.44e-2. Here, 6,819 genes are below this threshold.
- **Step 2:** We identify the 30th percentile for all 16,954 genes in the respective SNAT dataset for the mouse (i.e. 4 mSCs for mSCs application or 4 nmSCs for the nmSCs application).
  - In mouse mSCs, this 30% threshold is: 14.75. Here, 5113 genes are below this threshold (3,105 of these are also lowly-expressed in corresponding human data). In mouse nmSCs, this 30% threshold is: 25.25. Here, 5,096 genes are below this threshold (3,167 of these are also lowly-expressed in corresponding human data).
- **Step 3:** We remove any genes found below the threshold in Step 1 and/or Step 2 from our list of TFs (but will still retain those genes as TGs). This filtering is meant to remove TFs that have low gene expression levels in the training data and/or mouse data from the pool of potential candidate TFs (to guide us in our feature selection of TFs for the TG). Please note that we do not filter the list of TGs based on expression levels. In mSCs, we flag 8,827 genes with low expression. Any TFs in this list of lowly-expressed genes will be removed. In nmSCs, we flag 8,748 genes with low expression. Any TFs in this list of lowly-expressed genes will be removed.

Alzheimer's disease (AD):

Please note that we adapt our steps to be closely-related to those for Schwann cells (SCs).

- **Step 1:** We identify the 40th percentile for all 17,926 genes in the training gene expression dataset for that cell-type and/or condition (AD versus Controls).

- **Step 2:** We will remove any genes found below the threshold in Step 1 from our list of TFs (but will still retain those genes as TGs). This filtering is meant to remove TFs that have low expression levels in the training data and/or mouse data from the pool of potential candidate TFs (to guide us in our feature selection of TFs for the TG). Please note that we do not filter the list of TGs based on expression levels.

§B.1.7.2B Removing TFs based on relative gene expression levels for remaining TFs

After we perform **§B.1.7.2A**, we essentially have a reduced list of TFs from **§B.7.1C**, which meet a minimum percentile of relative gene expression levels in the cell-type. We note that the list of TFs to remove based on the criteria in **§B.1.7.2A** remains the same for each TG for the cell-type (since that filtering is at a global level for all genes in the (training) dataset). Then, we perform an additional filtering step to remove TFs in **§B.1.7.2A** that have lower expression levels relative to the other TFs in **§B.1.7.2A**. In this way, this is a more tailored filter applied based on the specific TFs that remain after **§B.1.7.2A**. We utilize the training gene expression data for all pooled TFs in **§B.1.7.2A** and determine the TFs that have relative gene expression above the 40th percentile (compared with the other pooled TFs). Thus, we determine the top 60th percentile of TFs (out of the pooled) that we should keep, filtering out the TFs from **§B.1.7.2A** to only include these TFs with relatively high expression**.**

§B.1.7.3 Pruning TFs down based on motif-score criteria

We also will use motif-score criteria to help us prune down the list of TFs for the TG.

§B.1.7.3A Identifying common TFs in both motif-based GRNs

We keep all the TFs found in common in **§B.1.7.1A** (scGRNom-based) and **§B.1.7.1B** (filtered monaLisa) as those are highly-confident TFs. That is, these TFs were found as viable candidate TFs for the TG using 2 separate methods; this instills confidence that these TFs could be key to consider.

§B.1.7.3B Filtering based on metrics for the monaLisa motif-based GRN

We filter the list of TFs found in **§B.1.7.2B** based on additional criteria, which we will mention in these upcoming steps. Since our monaLisa motif-based GRN (**§B.1.7.1A**) tends to be more extensive, we will apply filters to the TFs in **§B.1.7.2B** based on their relative performance for various motif-based metrics in **§B.1.7.1A.** In **§B.1.7.0B**, we identified a list of raw scATAC-seq peaks that are cell-type-specific. Then, in **§B.7.0C**, we identified peak-to-TG links (i.e. raw scATAC-seq peaks mapped to the TG via regulatory element relationships). Both approaches are different ways to annotate and glean information about the raw scATAC-seq peaks for the cell-type. For the given TG, there may be overlaps between the peaks in **§B.1.7.0B** with those in **§B.1.7.0C.** Here, we determine if the TG

contains any peaks (in §B.1.7.0C) that are also cell-type-specific (§B.1.7.0B). Then, we retain TFs based on various criteria.

- *If the TG has at least 1 cell-type-specific scATAC-seq peak:* If the TG does have scATAC-seq peaks that are cell-type specific (i.e. we identified cell-type-specific peaks for the TG), then we adjust our criteria to focus more on TFs that bind to the cell-type-specific scATAC-seq regulatory peaks as those TFs may help drive cell-type-specific gene expression levels. The minimum percentile for motif-matching scores that we use to filter out the TFs predicted by monaLisa is: *min_monaLisa_percentile* = 80% if the TG has cell-type-specific peaks.
- *If we found NO cell-type-specific scATAC-seq peak for the TG:* The minimum percentile for motif-matching scores that we use to filter out the TFs predicted by monaLisa is: *min_monaLisa_percentile* = 75% if we could not identify any cell-type-specific peaks for the TG.

All TFs need to fall in groups A, B, and C, which are global for the TG.

  - **A**: We find which TFs are bound to at least the *min_monaLisa_percentile* # of all regulatory regions (whether cell-type-specific or not) for that TG
  - **B**: Which TFs have a median motif-matching score that is at least *min_monaLisa_percentile* of that for the other TFs
  - **C**: The same as B, but focused on the average motif matching score

If the TG has cell-type specific peaks, we have additional criteria, for the genes to fall into (so they need to fall in groups A, B, C, D, E, and F, and we have a stricter *min_monaLisa_percentile*). The groups D to F are more tailored for the cell-type-specific regulatory regions for the TG.

  - **D**: We find which TFs are bound to at least the *min_monaLisa_percentile* # of only the cell-type-specific regulatory regions for that TG
  - **E**: Which TFs have a median motif-matching score for the cell-type-specific regulatory regions for the TG that is at least *min_monaLisa_percentile* of that for the other TFs that bind to cell-type-specific regulatory regions for the TG.
  - **F**: The same as E, but focused on the average motif matching score

Thus, §B.1.7.3B results in a set of TFs from monaLisa filtered down from that in §B.1.7.2B**.**

§B.1.7.3C Strong list of candidate TFs based on binding information

At this point, the TFs we identify for the TG thus are found in **§B.1.7.3A** (common TF in both scGRNom-based GRN and monaLisa-based GRN) and/or **§B.1.7.3B** (strong metrics in monaLisa-based GRN relative to other TFs in monaLisa-based GRN for the TG). We pool together the list of TFs from **§B.1.7.3A** and/or **§B.1.7.3B** as a strong list of potential candidate TFs based on binding information.

These steps so far represent our improvements in determining potential candidate TFs for the TG based on TF binding profiles. Our enhancements include:

- Predict TF binding for more regulatory interactions besides Distal (Enhancer) – Promoter – Gene
- Include potential "promoter-only" TGs that may lack data on regulatory interactions
- Use more comprehensive PWMs. Integrate more tools to infer TF binding from PWMs
- Utilize more flexible pruning criteria based on motif binding scores
- Retain TFs with a minimum average gene expression level in training data

§B.1.7.4 TF Augmentation Step:

This is our expansion step, where we build on the list of TFs that are found by **§B.1.7.3C**. For each of these TFs in **§B.1.7.3C**, we will additional TFs based on various TF similarity approaches and colocalization/cobinding measures that we have derived. For each TF, we prioritize adding TFs that not only have high scores for the corresponding metric but also contain known links in the PPI (if the TF is found in the PPI); if the TF forms a known complex with other TFs, we especially focus only on the TFs that it forms a complex with and see which of those TFs have high values for that given metric. We focus on adding in TFs that are not found in **§B.1.7.3C**, so that unique TFs can be added in this step; nonetheless, it is possible that a given TF may be added in multiple times in this Step. We use these 3 approaches to identify highly-similar TFs to add to the list of TFs from our final motif-based GRNs.

- **§B.1.7.4A** *Colocalization Approach 1*: we add TFs with the highest colocalization score based on monaLisa-based derivations (high Jaccard Similarity values).
- **§B.1.7.4B** *Colocalization Approach 2:* we add TFs with the highest colocalization score based on scGRNom-based derivations (high Jaccard Similarity values).
- **§B.1.7.4C** *Gene Ontology Molecular Function Similarity*: add TFs with the highest similarity with that TF that are also found in the PPI.

§B.1.7.4A Determining Colocalization Approach 1:

In approach 1, we will focus on cell-type-specific scATAC-seq peaks. We utilize these cell-type-specific raw scATAC-seq peaks (from **§B.1.7.0**) for our *Colocalization among TFs* analysis approach 1). Thus, we identify scATAC-seq peaks unique for that cell-type, ensuring no significant overlap with peaks from other cell-types (Zhang et al. 2021). Then, we run the monaLisa tool to analyze the colocalization among TFs. We use method = "matchPWM" and min.score of 7.5 to identify more TFs binding to the regulatory region for each of the cell-type-specific scATAC-seq regions that we determined. We use our more comprehensive input Position Weighted Matrix (PWM) of 8,903 PWMs.

Using monaLisa (Machlab et al. 2022), we predict TF binding within these specific peaks, noting the TFs and their binding locations. The monaLisa tool returns TFs predicted to strongly bind to each given regulatory region along with their scores and predicted positions within the given regulatory region. To determine potential colocalization among a pair of TFs for each region, we determine if there is any configuration of binding positions where the 2 TFs may not overlap (or at least bind 1 base pair apart from each other). We assign a score of 1 if there exists such a configuration between those TFs for that regulatory region, else a score of 0. Then, we aggregate the number of regulatory regions for which these 2 TFs have score of 1 to determine the number of regions with colocalization among the TFs.

We note that TFs that may overlap could form complexes. Nonetheless, the TF augmentation step already adds TFs that form known complexes with each other to the list of proteins. Hence, for this approach, we will just assume that TFs that are overlapping across all possible binding configurations for the regulatory region may compete for binding to that region and likely do not exhibit cooperative behavior at that region; and with that prediction, both TFs receive a 0 score for that region. Thus, for each TF pair, a score of 1 is assigned per peak if they have non-overlapping potential TFBSs, suggesting cooperativity. Aggregate scores for all TF pairs are calculated across the peaks. Finally, we scale these pairs using a Jaccard Similarity (JS) index, with higher scores indicating greater colocalization. Here, colocalization is 2 TFs that bind within the same enhancer but not to the same motif. $JS(TF_i, TF_j) = \frac{\text{\# of peaks with potential non-overlap in binding for } TF_i \text{ and } TF_j}{\text{\# of peaks bound by } TF_i \frac{\text{and}}{\text{or}} TF_j}$. For each TF in **§B.1.7.3C**, we add TFs with the highest colocalization score based on monaLisa derivations (high JS values).

§B.1.7.4B Determining Colocalization Approach 2:

In Approach 2, we will focus on all the raw scATAC-seq peaks (including the cell-type-specific, mapped, and unmapped peaks). For analyzing colocalization among TFs (Approach 2 in the main paper for *Colocalization among TFs*), we run scGRNom (Jin et al. 2021) Step 2 (with our adaptations) on all the raw single-cell ATAC-seq regions for that given cell-type to predict TFs with strong binding. We run motifmatchr (Schep et. al 2023) on all raw scATAC-seq peaks and identify TFs with strong binding for each peak. Then, we determine the JS of binding among TFs by analyzing the # of common regulatory regions (out of all raw regions in the cell-type) they share strong binding for. Next, we calculate JS between any 2 TFs as: $JS(TF_i, TF_j) = \frac{\#\text{ of peaks bound by } TF_i \text{ and } TF_j}{\#\text{ of peaks bound by } TF_i \frac{\text{and}}{\text{or}} TF_j}$. Then, we consider TFs with a high JS score as being very similar to each other (e.g. potentially sharing same TF family of motifs) based on this derived predicted TF-DNA-binding data. For each TF in **§B.1.7.3C**, we add TFs with the highest colocalization score based on scGRNom-based derivations (high JS values).

§B.1.7.4C Determining Colocalization Approach 4:

We employ clusterProfiler (Wu et al. 2021) to determine TF similarity based on Molecular Function (MF). Using the godata() function on the org.Hs.eg.db database (Carlson et. al 2019), we generate semantic similarity data for Gene Ontology (GO) Molecular Function (GOMF). We then apply the mgeneSim() function with default settings, utilizing the Wang measure and BMA (Best-Match Average) approach. This method helps infer genetic interactions and PPIs, as closely-related TFs often exhibit similar behavior (Yu). For each TF in **§B.1.7.3C**, we add TFs with the highest similarity with that TF that are also found in the PPI network.

§B.1.7.4D Combining TFs from 3 Approaches:

For each TF in **§B.1.7.3C**, we add a list of TFs based on **§B.1.7.4A-C**. We do this for all of the TFs in **§B.1.7.3C**. Ultimately, this results in new TFs (some may overlap with those in **§B.1.7.3C**), which we add to the list of TFs in **§B.1.7.3C** to retrieve our updated list of TFs.

§B.1.7.4E Adding TF complexes:

Here, we add TFs that are found to form physical complexes with any of the TFs in **§B.1.7.4D**. Then, we use data on known TF complexes (i.e. TF pairs with weight 1 in comprehensive input PPIN). While we had done this in **§B.1.7.1C**, we recognize that some of these TFs may have been eliminated during the pruning steps in **§B.1.7.2** and **§B.1.7.3**. This is our final list of *N* candidate TFs for the given TG.

Section §B.1.8 Building out Protein-Protein Interaction (PPI) Networks (PPINs)

We assume users typically have a PPIN available for their specific needs that they can input to NetREm along with their respective gene expression data (and optional prior GRN knowledge). The proteins are represented by their HGNC (HUGO Gene Nomenclature Committee) gene symbols. PPI network edges are undirected (i.e. only 1 edge weight for each TF pair) with non-negative weights (i.e. weights $w > 0$). For both humans and mice, we utilize the STRING PPI database (Szklarczyk et al. 2023). STRING has average combined scores (across several evidence types), which we scale to be between 0 and 1. This STRINGdb database focuses on proteins that have a function association and specific partnership. These proteins are physically associated with each other or are indirectly associated (e.g. work together in a common metabolic or signaling pathway, regulate each other through intermediaries, contribute to a common biological cause)(Szklarczyk et al. 2023). That is, according to BioGRID (Oughtred et al. 2021) (one of the PPI resources used by STRING) the definition of interaction among 2 proteins can include: direct physical binding of 2 proteins, co-existence in a stable complex, and/or genetic interaction. Thus, this STRING PPI database has edges that represent protein-protein associations among proteins that jointly contribute to a shared, specific, and meaningful function, regardless of whether the proteins physically bind to each other. Currently, the latest STRINGdb release is version 12 (v12), which includes data since July 26, 2023.

Our final comprehensive PPI networks $\mathbb{W}$ for each organism contain PPI links with edge weight $w > 0.01$ and are thereby typically partly-connected networks. There are no self-loops in our network. Then, we filter these networks to get our TF-TF PPI networks. For a given TG in a context, we will subset this network $\mathbb{W}$ for just its N candidate TFs, to get a TF-TF PPIN. Here, the maximum weight for any TF_i- TF_j interaction is 1 (that is, $0.01 < w_{ij} \leq 1$), where $TF_i \neq TF_j$. We add artificial links ($\eta = 0.01$) to this network to make it connected pairwise for all N candidate TFs to yield W, our TG-specific TF-TF PPI network: $0.01 \leq w_{ij} \leq 1$, where $TF_i \neq TF_j$ and all N candidate TFs have pairwise edges with the other $N - 1$ edges. We do not consider self-loops (i.e. we do not consider TF_i- TF_i interactions of the same TF with itself). Instead, we set $W_{ii} = \frac{d_i}{N-1}$ where TF_i's degree (connectivity) with $N - 1$ other TFs is $d_i = \sum_{k \neq i} w_{ik} > 0$. Below, we explain the processes for obtaining these networks in humans and in mice.

§B.1.8.1 Final Comprehensive PPI Networks (PPINs)

Organism: Humans (Homo sapiens)

We construct a comprehensive human PPI network of associated interactions, with weights $w > 0$, from multiple sources. STRINGdb v12 (Szklarczyk et al. 2023) provides 6,857,702 undirected edges for humans (id: 9606). Hu.MAP 2.0 (Drew et al. 2021) contributes ≈8.75 million edges with probability weights, which we use as weights. Other data, including protein complexes and PPIs, comes from MaayanLab (e.g. BioGRID, DIP, HPRD, HubProteins, NURSA, CORUM) (Giurgiu et al. 2019), Human Reference Interactions (HuRI) (Luck et al. 2020), TF complexes in: PWMs, Contextual PPI (Kotlyar et. al 2022), KEGG Path Networks (Kanehisa and Goto 2000) we web-scraped. These sources are integrated to build our final PPIN. A weight of 1 is assigned to any TFs found to physically interact in complexes with each other in any source. For PPI links found in multiple sources, we use the highest weight for that link. We filter this network to only include links with edge $w > 0.01$. Ultimately, we have 8,701,636-undirected PPI edges in our final comprehensive PPIN $\mathbb{W}$, our ground truth for humans. All weights are between 0 (missing link) and 1 as this network is partly-connected. Here, links may correspond to direct and/or indirect physical TF-TF interactions with evidence of functional association in humans. We may refer to this network as *NewNet* and will use this for many applications including our Schwann cell application and AD versus control application (i.e. Applications 6 to 7).

V11 network in humans:

The human STRINGdb version 11 (v11 or V11) network refers to the PPI network with updates from January 19, 2019, to October 17, 2020: 11,759,454 rows of [Protein1 ID, Protein2 ID, combined score]. We map these protein IDs to gene symbol names using the more updated v12 protein information file for 21,840 proteins in humans. In the process, we obtain our v11 network that has [human gene 1, human gene 2, combined score]. We average combined scores across the same human gene pairs (in case there are duplicate pairs with different scores) to obtain the average combined scores. Ultimately, the final comprehensive PPI v11 network for humans has 9,521,964 rows corresponding to 4,760,982 unique undirected edges among proteins. This human PPI network $\mathbb{W}$ reflects the older (potentially outdated) knowledge regarding PPIs among the proteins in humans and will be a baseline.

Organism: Mouse (Mus Musculus)

We utilize 2 different versions (v#) of the STRING PPI database for mice (id: 10090).

- *Older PPI network in mice (v11 network):*

The mouse STRINGdb version 11 (v11) network refers to the PPI network for mice as of January 19, 2019, to October 17, 2020, which has 11,944,806 rows of [Protein1 ID, Protein2 ID, combined score]. We map protein IDs to

gene symbol names using the more updated v12 protein information file for 21,840 proteins in mice. In the process, we retain 8,600,490 rows for our v11 PPIN that has [mouse gene 1, mouse gene 2, combined score]. Then, we average combined scores across the same mouse gene pairs (in case there are duplicate pairs with different scores) to obtain average combined scores. Ultimately, the final comprehensive v11 PPIN $\mathbb{W}$ for mice has 8,600,490 rows corresponding to 4,300,245 unique undirected edges among proteins. This mouse PPI network reflects the older (potentially outdated) knowledge regarding PPIs among the proteins in the newer v12 PPIN and will be a baseline.

- *Newer PPI network in mice (v12 network):*

We perform the same data pre-processing steps as we did for the v11 network for the mouse STRINGdb v12 network that has 12,684,354 rows. Then, we use motif complexes from the JASPAR2022 (Castro-Mondragon et al. 2022) database for mice to adjust the PPI scores to be 1 (our maximum) for 11 TF pairs known to be in a complex (i.e. AHR::ARNT, HAND1::TCF3, GATA1::TAL1, POU5F1::SOX2, PPARG::RXRA, FOS::JUN, ARNT::HIF1A, NR1H3::RXRA, STAT5A::STAT5B, BACH1::MAFK, ZIC1::ZIC2). Ultimately, the final comprehensive PPI v12 network for mice $\mathbb{W}$ has 12,684,354 rows corresponding to 6,342,177 unique undirected edges among the proteins. This is the newer PPI network for mice, which reflects the current knowledge regarding PPIs among the proteins.

§B.1.8.2 Fully-Connected Comprehensive TF-TF PPINs

TFs are types of proteins. Thus TF-TF interactions are a subset of the PPIs that are measured. Hence, in our real-world applications, we will filter the final comprehensive PPIN for the respective organism to include only PPI links among the relevant TFs. That is, the comprehensive TF-TF PPINs are subnetworks of final comprehensive PPINs, and they refer to direct and/or indirect TF-TF interactions with existing prior knowledge predictions to support them (i.e. $w > 0.01$). When we run NetREm, we assign a default edge weight of $\eta = 0.01$ to any TF-TF edges that are missing in the TF-TF PPIN for the organism (i.e. every TF has a weighted pairwise link to every TF whether that link is known or potentially artificial); we add these artificial TF-TF links to help ensure a fully-connected PPI network for NetREm's regularized regression (for numerical stability and more accurate results) and guide the discovery of novel TF-TF coordination links. That is, we run NetREm on our final fully-connected comprehensive TF-TF PPIN where all TFs have pairwise links to each other with weight $\eta = 0.01 < w \leq 1$.

Organism: Humans (Homo sapiens)

We apply the final comprehensive fully-connected human TF-TF PPI networks to predict TF-TF coordination networks and TF-TG regulatory networks for our human applications (#1, 2, 5, 6, 7). For the Schwann cell and AD

applications (#6 and 7, respectively), we incorporate prior GRN knowledge from multiomics data, such that the number and set of N candidate TFs may differ across TGs in each cell-type. Thus, we only provide main metrics related to fully-connected TF-TF networks for our hESCs and HSCs applications (#1 and 2, respectively), which we run NetREm for in the absence of prior knowledge.

In hESCs, we have the same $N = \mathcal{N} = 207$ (or $N = 206$ if TF is a TG) candidate TFs fixed for each TG. For 1,235 of 1,250 TGs where the TG is not a TF (i.e. $N = 207$), we have $\frac{(207)(206)}{2} = 21{,}321$ TF-TF edges (corresponding to 42,642 rows) in our fully-connected TF-TF PPIN, comprising 10,777 known and 10,544 artificial edges. Similarly, in HSCs, when the TG is not a TF (i.e. $N = \mathcal{N} = 178$), we have 15,753 TF-TF edges in our fully-connected TF-TF PPI network, comprising 8,298 known and 7,455 artificial edges.

Organism: Mouse (Mus Musculus)

We apply the final comprehensive fully-connected mouse TF-TF PPINs to predict TF-TF coordination networks and TF-TG regulatory networks in mouse ESCs, where we have the same $N = \mathcal{N} = 195$ (or $N = 194$ if the TF is a TG) candidate TFs fixed for each TG. In the cases where the TG is not a TF (i.e. $N = 195$), we will have 18,915 TF-TF edges (corresponding to $(\mathcal{N})(\mathcal{N} - 1) = 37{,}830$ rows) in our TF-TF PPIN, comprising known (v11: 5,044; v12: 6,856) and artificial (v11: 13,871; v12: 12,059) TF-TF edges. Both networks lack any known interactions for 2 orphaned and singleton TFs (AES and STRA13) with other TFs. There are 749 out of 5,044 edges in the v11 network that are not found in the v12 release and may have been removed as potential False Positives. The removal of edges from v11 to v12 networks may be due to many factors. As new experimental data becomes available and existing data is re-evaluated, previously identified PPIs might be reclassified. Some interactions may have been based on data that, upon further scrutiny, is now found to be inaccurate or not reproducible, leading to their removal; moreover, PPI databases may update their criteria for including PPIs to reflect the latest in bioinformatics research and protein interaction validation methodologies. This might involve higher thresholds for confidence scores, leading to the exclusion of lower-confidence interactions that were included in previous versions. Similarly, 2,561 out of the 6,856 edges in the v12 network are relatively recent discoveries as they are not in the v11 network. In fact, there are 25 singleton TFs in the v11 network that have known interactions in the v12 network.

In mDCs, we have $N = N^* = 93$ (or $N = 92$ if the TF is a TG) candidate TFs fixed for each TG. In the cases where the TG is not a TF (i.e. $N = 93$), we will have 4,278 TF-TF edges (corresponding to 8,556 rows) in our fully-connected TF-TF PPIN, comprising known (v11: 863; v12: 3,110) and artificial (v11: 3,415; v12: 1,168) TF-

TF edges. We run NetREm using the older v11 TF-TF PPIN to predict TF-TF coordination scores and gauge how effective the predictions are in terms of predicting future knowledge (i.e. links known in v12 TF-TF PPIN).

Section §B.1.9 NetREm Parameters for Applications

By default, NetREm reserves 30% of samples for testing and utilizes the remaining 70% for training. This is a partitioning used by other methods (Jin et al. 2021). Here, β is the network-constrained hyperparameter and comes beforehand and is used to define network regression gene embeddings $\tilde{X}$ and $\tilde{y}$; α is the sparsity prior hyperparameter (and optimizes prediction of $\tilde{y}$ from $\tilde{X}$). Thus, we often select β first and then may use LassoCV solvers to optimize α based on $\tilde{X}$ and $\tilde{y}$. Please note that for all applications (i.e. 1 to 7), we use our preprocessed respective TF-TF PPI network (PPIN) with positive weights as a network regression constraint; we do not fit a y-intercept for our NetREm or our baseline models. We run NetREm each time (1 by 1) for each TG in the cell-type (and/or context) to yield TG-specific TF-TF coordination network B and N^* final TFs for the TG (where N^* can vary across TGs even for fixed N candidate TFs since N^* TFs are those with $c^* \neq 0$). We then aggregate our results to yield our cell-type TF-TG regulatory network (comprising individual TF-TG links) and our cell-type TF-TF coordination network $\bar{B}$.

For our benchmarking applications, please note the following NetREm hyperparameters with more details in **Table B.17** regarding the # of TGs, # of cells M, fixed set of N candidate TFs for TG (held constant across all TGs for the $\mathcal{N}$ potential cell-type TFs; if the TG is also one of the $\mathcal{N}$ TFs, we just remove that self-TF (corresponding to the same TG) from the list of candidate TFs so there is no overlap between the candidate TFs and the TG → $N = \mathcal{N} - 1$ candidate TFs for the TG). **Human (hESCs) Embryonic Stem Cells (ESCs):** we use $\beta = 1$ and then vary α for each of our 6 SERGIO-simulated gene expression datasets (varying noise parameter: {30%, 60%, 90%}, varying training cells M: {70; 700}). We may use LassoCV to optimize α based on 5-fold cross validation (CV). We also try out 5 fixed values of α: 0.01, 0.025, 0.05, 0.075, 0.1. Thus, the following 6 NetREm runs (for each of the 1,250 TGs in each of the 6 gene expression datasets) are: NetREm (β = 1, α = LassoCV), NetREm (β = 1, α = 0.01), NetREm (β = 1, α = 0.025), NetREm (β = 1, α = 0.05), NetREm (β = 1, α = 0.075), NetREm (β = 1, α = 0.1). **Mouse (mESCs) ESCs**: we use a fixed $\beta = 1$ and α = 0.05. We mainly alter the input STRINGdb PPIN that is used for our network constraint and run NetREm with the older Version 11 network (updated between January 2019 and October 2020) and the current newest V12 PPIN (updates since July 2023).

That is, we have 2 runs of NetREm for each of the 19,225 TGs: NetREm (β = 1, α = 0.05) with the older V11 STRING PPI network and NetREm (β = 1, α = 0.05) with the newer V12 STRING PPI network. **Mouse Dendritic cells (mDCs):** we vary α for 8 values ($\alpha \in \{0.01; 0.025; 0.05; 0.075; 0.1; 0.125; 0.15, \text{LassoCV}\}$) for a fixed $\beta = 1$. We do this for older PPIN V11 and newer PPIN V12. Thus, these are the following 16 runs that we do for each of the 9,087 TGs for V11 and V12 networks; $\text{NetREm}\ (\beta = 1, \alpha = 0.01), \text{NetREm}\ (\beta = 1, \alpha = 0.025), \text{NetREm}\ (\beta = 1, \alpha = 0.05), \text{NetREm}\ (\beta = 1, \alpha = 0.075), \text{NetREm}\ (\beta = 1, \alpha = 0.1), \text{NetREm}\ (\beta = 1, \alpha = 0.125), \text{NetREm}\ (\beta = 1, \alpha = 0.15), \text{NetREm}\ (\beta = 1, \alpha =$ from LassoCV).We also alter the network-regularization penalty β so $\beta \in \{0.1;\ 10\}$ and optimize for α using LassoCV for the newer PPIN V12. Please note we already used LassoCV for $\beta = 1$. Thus, these are the 2 runs we do for the V12 networks: NetREm (β = 0.1, α = from LassoCV), NetREm (β = 10, α = from LassoCV). **Human Peripheral Blood Mononuclear Cells (PBMCs):** We run NetREm (for each of the 13,714 TGs) with $\beta = 1$ and $\alpha = 0.1$ for each of the 9 immune cell sub-types (with Cluster #s): Naïve CD4 T (C0), CD14 Monocytes (C1), Memory CD4 T (C2), B-cells (C3), CD8 T (C4), FCGR3A Monocytes (C5), Natural Killer (C6), Dendritic Cell (C7), Platelet (C8). Each of these 9 cell-types has 13,714 TGs. We do this for the comprehensive human input PPIN (we use for SCs and AD applications 6 and 7 respectively; we call this *NewNet* elsewhere) that has V12 human PPIs and beyond (from various sources detailed in §B.8). We also do this for the older STRINGdb V11 PPIN. So, there are 18 different runs (2 for each cell sub-type based on input PPIN used). **Human Hematopoietic Stem Cells (HSCs):** We have 4 different NetREm runs, which we do on NewNet (comprehensive input human PPIN): NetREm (β = 1, α = 0.01), NetREm (β = 0.5, α = 0.01), NetREm (β = 1, α from LassoCV), NetREm (β = 10, α from LassoCV). **NetREm comparisons with RTNduals:** we run NetREm with $\beta = 1$ and use LassoCV to optimize α for 13 different applications where RTNduals (Chagas et al. 2019) returns outputs for TF-TF coordination links (please see §B.10.9). We run NetREm using NewNet (comprehensive input human PPIN) and the older V11 human PPIN for these 13 applications. These applications are 8 Control gene expression data (Mathys et. al 2019) for neurons/glial cells (also used in AD application but for RTNDuals comparison we fix the set of N candidate cell-type TFs and use all 17,926 TGs): Endothelial Blood Brain Barrier (BBB) cells, Oligodendrocytes, Oligodendrocyte Precursor Cells (OPCs), Excitatory Neurons (ExNs), Inhibitory Neurons (InNs), Pericytes, Astrocytes, Microglia. We also use (Lake et al. 2018) that has control gene expression for 15,693 TGs in 4 cell types: Microglia, InNs, ExNs, Oligodendrocytes. Lastly, we also use the GTEx version 8 Schwann cell (SC) pooled data that has 14,144 TGs.

For our applications in SCs and AD versus controls, we use $\beta = 100$. We use the comprehensive human PPIN as our network regression constraint. For mSCs, we use a stricter sparsity hyperparameter of $\alpha = 0.1$. In nmSCs, we used a relatively more relaxed hyperparameter of $\alpha = 0.025$. Further, we filter predictors based on goodness of fit, retaining only TGs with test NMSE values less than 1.01 in nmSCs. We also filter TGs based on test MSE values where we impose a harsher filter on nmSCs due to the relaxed sparsity during model training. That is, we retain mSC TGs where test MSE values are less than 1.5 and retain nmSC TGs where test MSE values are less than 1.0005. For AD applications we fix $\beta = 100$, exclude the y-intercept. For other 8 neuronal/glial cell-types in AD and/or control, we use Lasso with α = 0.25. For ease of comparison across 18 TF-TG regulatory networks and respective TF-TF interaction networks, we apply no additional filters. **Table B.18** provides metrics on the # of TGs, # of cells for training NetREm models (M), and metrics on values of N, the # of candidate TF predictors. The value of N may vary across the TGs because of the prior reference GRN of TF-RE-TG links (derived from multi-omics data, which we explain in **§B.1.7**) where RE are Regulatory Elements (REs) linked to the TG by peak-to-TG links.

Section §B.1.10 Evaluation

For the first 4 applications, we have prior ground truth GRN validation data. In hESCs, we have 5,050 TF-TG signed ground truth GRN links subset from the original atlas of regulatory links (Sharov et al. 2022) for 1,250 randomly-selected TGs. In HSCs and mESCs, we have pooled gold standard GRN networks from (Zhang et al. 2023). We measure the performance of our NetREm models in various ways. These include running benchmark regression models (BRMs, e.g. LassoCV, ElasticNetCV, RidgeCV, LinearRegression (Pedregosa et al. 2011), GRNBoost2 (Moerman et al. 2019)) on the same data. We use SNPs related to disease/traits in GWAS and/or to changes in expression via eQTL analysis. These SNPs are mapped onto our TF-TG regulatory networks (comprise TF-RE-TG links) using impacted TFBSs and REs for TGs. Further, we perform gene enrichment analysis of TGs and relative gene percentile analysis. We have experimental data for 8 core TFs (EGR2, NR2F2, RXRG, SOX10, SREBF1, STAT1, TEAD1, YY1) in SCs. In addition, we utilize a CPPID with annotated TF-TF links (Kotlyar et. al 2022) associated with various diseases/traits. We also benchmark NetREm's performance in applications 1 to 5.

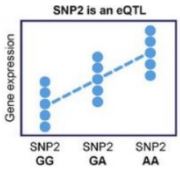
§B.1.10.1 Benchmark Regression Models (BRMs): Baseline Models (Expression-based Metrics for TGs)

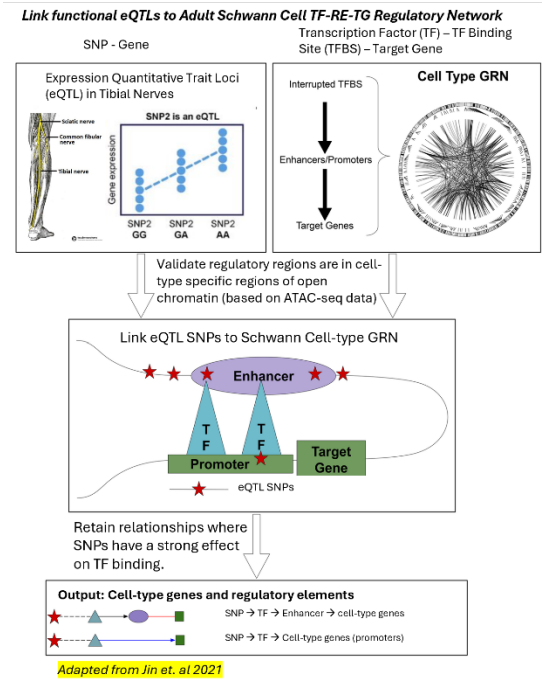
Model selection, regularization techniques, and ensemble methods are commonly used to navigate the bias-variance tradeoff, aiming to find a model with an optimal balance between bias and variance. To assess the performance of our NetREm model, we employ a comparative approach involving 4 benchmark regression models (BRMs) as a

baseline: ElasticNetCV, LassoCV, Linear Regression, and RidgeCV, all implemented using Python's scikit-learn (Pedregosa et al. 2011) package. For each TG within a specific cell type, we fit NetREm and these benchmark models for the same N candidate TF predictors (where N may vary in number and specific TFs selected across TGs if prior knowledge is incorporated and/or the TG is also a cell-type TF). All 5 models fit a y-intercept. These 4 BRMs enable us to gain a better understanding of NetREm's performance. All models, except for Linear Regression, undergo 5-fold cross-validation (CV) to optimize hyperparameters. Statistical significance of performance differences between NetREm and benchmarks is assessed using 1-sided t-tests of whether NetREm consistently yields lower test MSEs.

B.1.10.2 Expression quantitative trait loci (eQTL)-based validation

We use eQTL data, which links single nucleotide polymorphisms (SNPs) to changes in the expression of various TGs. Integrating eQTL data may help us verify some of the TF-Regulatory Element-TG links inferred by NetREm. For AD we incorporate additional eQTLs for 4 of our cell-types from Zeng et al. (Zeng et al. 2022); we did not use this source when defining eQTL-based regulatory regions. For SCs, we recognize that the Tibial Nerve eQTL dataset is limited in terms of its associations as it only has sequencing for 532 human samples, and studies observe that relatively small sample sizes for eQTL data may likely result in relatively low statistically significant SNP-TG associations(Ndungu et al. 2020); moreover, this data (i.e. left peripheral tibial nerve from the gastrocnemius region) is a proxy for SCs, given the current unavailability of a SC-specific cell-type eQTL dataset. Hence, for eQTL validation for SCs, we allowed weak effects (along with strong effects) for motifbreakR (Coetzee et al. 2015) predictions of SNP-TF links. For AD applications, we mandate that SNP-TF links ought to have strong effects.





To understand the relationship between SNPs, TFs, and TGs, we employ a computational pipeline that begins with NetREm for the initial inference of TF-Regulatory Element (RE)-TG links. That is, for our TGs, we may overlay our final cell-type TF-TG regulatory networks (predicted by NetREm) with the TF-RE-TG links (based on the prior GRNs); this is since the prior GRNs do initial feature selection to identify N candidate TFs for the given

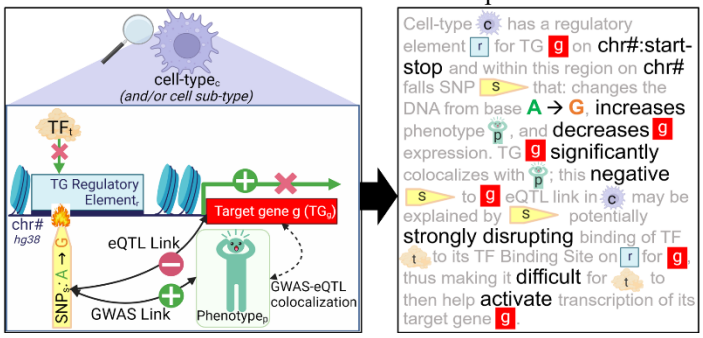
TG out of the $\mathcal{N}$ potential cell-type TFs overall (where some of these N candidate TFs have TF-RE-TG links based on the motif-based GRN and other candidate TFs are augmented with indirect association with these TFs). Then, NetREm selects $N^* \leq N$ optimal TFs with $c^* \neq 0$ for the TG; thus we may see which of these N^* TFs has TF-RE-TG links for the TG and use that information to retrieve our TF-RE-TG links (for other TFs that lack strong binding support, we tend to assume they may have more indirect binding mechanisms and do not focus on those). To validate these inferred relationships, we use motifbreakR tool designed to predict the impact of SNPs on TF binding to their binding sites (TFBSs). This allows us to assess whether SNPs within the REs can explain the observed cell-type-specific eQTL data. Specifically, we examine SNPs that not only fall within the regulatory elements of TGs but also affect TF binding and correlate with changes in TG expression. Please note that the image on the right is adapted from (Jin et al. 2021).

SNPs mapped on original TF-TG regulatory networks

Nonetheless, in the absence of prior reference GRNs for the TGs, we may still use NetREm's final TF-TG regulatory networks where for a given TF-TG regulatory link, we can see if: the eSNP-TF exists (based on motifbreakR) and the eSNP-eTG link exists (based on expression eQTL eSNP-eTG links). This may yield more results (than overlaying with TF-RE-TG links can).

GWAS-eQTL colocalization analysis

We incorporate Genome-Wide Association Studies (GWAS) to further enrich our analysis. GWAS data associates SNPs and various traits, including diseases. In this way, we may predict pathways from non-coding SNPs to disease phenotypes, via TFs with altered TF binding to TFBSs, impacting TF interactions with each other and ultimately cascading to changes in TG expression. We specifically investigate SNPs that are part of our eQTL-based validation and associated with GWAS traits. To assess the statistical significance of links that influence both TG expression and GWAS traits, we perform GWAS-eQTL colocalization analyses using the coloc (Wallace 2020) tool for a given shared eQTL and GWAS SNP. We set posterior probability (PP4) threshold > 50% for statistical significance. Here, PP4 is the probability that the given SNP is associated and affects both traits: GWAS trait and TG expression trait (i.e. eQTL) (Giambartolomei et. al, 2014). We assess whether any statistically significant links exist that influence both TG expression and the given GWAS trait. Specifically, we consider PP4 $\geq$ 70% as a threshold for statistical significance.



Adding Linkage Disequilibrium (LD) Analysis

To deepen our understanding of SNP co-inheritance patterns within the population (and their non-random association with each other), we conduct Linkage Disequilibrium (LD) analyses via LDlink (Machiela et. al 2015). This helps us correlate a pair of SNPs on the same chromosome. SNPs that are in LD are significantly correlated.

§B.1.10.3 Gene Enrichment Analysis

We use Gene Enrichment Analysis tools (e.g. MaayanLab (Rouillard et al. 2016), clusterProfiler (Wu et al. 2021), BaderLab (https://baderlab.org/), Metascape (Zhou et al. 2019b), and g:Profiler (Reimand et al. 2007) with $-\log_{10}(q)$ to validate biological relevance of our terms for our input lists of TGs (where q refers to a p-value (p) adjusted for the False Discovery Rate due to multiple hypothesis testing).

§B.1.10.4 Relative gene expression percentile analysis

Schwann cells (SCs)

We assess the relative gene expression levels of all predicted TGs for all of the TFs in mSCs and/or nmSCs in SCs using GTEx gene expression data across various tissues (Eraslan et. al 2022) We compare these results with the corresponding respective levels using expression data for 4 CNS cell types sourced from Lake et al. (Lake et al. 2018). While we perform this analysis for all TFs in mSCs and nmSCs, we are primarily interested in ascertaining whether the TGs for core SC TFs (e.g. EGR2, NR2F2, SOX10, SREBF1, TEAD1, YY1) exhibit cell-type specificity. To refine our exploration, we delve into comparative analyses of relative gene expression percentiles across 5 unique cell types in the nervous system, leveraging gene expression data from the aforementioned validation datasets. Recognizing variations in expression values and the spectrum of genes between datasets from (Lake et al. 2018) and (Eraslan et. al 2022), relative gene expression percentiles are employed as a common ground for comparison. Every gene within a given cell type and dataset is ranked by its expression level relative to its peers in that cell type, compiling these ranks for all TGs per TF in SCs, mindful of potential absences of some TGs from the datasets. This procedure is mirrored for 4 other CNS cell types, with an acknowledgment of possible missing TGs. To ensure robustness of our findings, we utilize 1-sided paired t-tests to compare the relative gene expression percentiles of TGs in SCs against the 4 CNS cell types, with non-TGs for the given TF as a baseline control. We retrieve p-values for all TFs in nmSCs (for their respective nmSC TGs) and for all TFs in mSCs (for their respective mSC TGs) and perform multiple-testing correction using the Benjamin-Hochberg (B-H) procedure. We retain significant comparisons as those where the adjusted p-value is less than 5%. Results are deemed meaningful if relative gene expression percentiles for our TGs are not only statistically higher in SCs compared to other CNS cell types but also exhibit specificity to this group of TGs (i.e., significance is not observed in control genes).

§B.1.10.5 Experimental data for Schwann cells (SCs)

We gathered as well as performed experimental analysis for TFs that play key roles in Schwann cells, such as: EGR2, TEAD1, SOX10, RXRG, SREBF1, STAT1, YY1, and NR2F2. Thus, we primarily analyze NetREm's predicted TF-TG regulatory links and TF-TF interactions with respect to these TFs. SOX10 has been shown to play key roles in mSCs and nmSCs (Balakrishnan et al. 2021) and EGR2 (may serve as either an activator or repressor of TG expression [18]) is important for mSCs (Svaren and Meijer 2008). We used biomaRt(Durinck et al. 2009) and g:Profiler(Reimand et al. 2007) to map rat/mouse gene names to the orthologous gene names in humans.

We consider *loss-of-function (LOF) TGs* as those that are impacted directly by the TF being knocked out. Then, we consider validated direct TGs as *LOF TGs* that have a ChIP-seq binding peak for that respective TF that is within 100 kilobases on the DNA; thus, *validated direct TGs* not only have knockout support (TG expression is associated with the presence or absence of that TF) but also strong evidence of TF binding nearby. Hence, we are more confident about the *validated direct TGs* but some of the *LOF TGs* could be indirectly regulated by a given TF. This additional data serves to both validate our predicted TF-TG regulatory links and to assess the strength of potential TF-TF interactions.

Target Genes (TGs) for core SC TFs Predicted by TF-TG regulatory links	Validation Data for Schwann Cells (SCs)	
	Mice/Rat Knockout	***Rodent ChIP-seq binding peak***
	TG expression affected by knockout of TF	TGs have a peak for that TF within 100 kilobases on DNA
Loss-of-Function TGs	✓	**?**
Direct TGs (Verified)	✓	✓
Other Candidate TGs	**?**	**?**

Data utilized: We have ChIP-seq (Chromatin Immunoprecipitation Sequencing) binding data from sciatic nerve of P15 rats (rn5 genome) for the following TFs: EGR2, SOX10 (Lopez-Anido et al. 2015; Srinivasan et al. 2012), and SREBF1 (manuscript in preparation). EGR2 and SOX10 are expressed exclusively in Schwann cells of peripheral nerve, and ChIP-seq raw data have been deposited in NCBI (Lopez-Anido et. al 2015). We obtain ChIP-seq data for STAT1 in differentiating rat SCs from NCBI GSE211337 for the rn6 rat genome assembly (Xu et al. 2023). For analysis of binding to known enhancers, we have employed H3K27ac ChIP-seq data from rat peripheral nerve (Lopez-Anido et al. 2015; Hung et al. 2015). We utilize ChIPseeker (Yu et al. 2015) to annotate peaks with TF

binding to the nearest gene, intergenic region, promoter, Transcription Start Site (TSS), Transposable Elements (TEs), exon, or intron. We also use rtracklayer (Lawrence et al. 2009) and the rn5tohg38.over.chain (Karolchik et al. 2009) file from UCSC (Karolchik et al. 2009) to perform liftover (2021) from rn5 to hg38 genomic coordinates for the ChIP-seq tracks (we similarly use rn6torn5.over.chain to map STAT1 TF ChIP-seq files to rn5, which may explain its relatively lower overlaps). For YY1, RXRG, NR2F2, TEAD1, we perform Cut&Run (Cleavage Under Targets & Release Using Nuclease) assays in the rat S16 Schwann cell (SC) line (GSE247955).

For loss-of-function (LOF) data, we have obtained expression profiling data for sciatic nerve of knockout mice for EGR2(Le et. al 2005), and Schwann cell-specific knockout of YY1 (He et. al, 2010). For TEAD1, we employ RNA-seq data from analysis of a Schwann cell-specific knockout of YAP/TAZ coactivators of TEAD TFs (Poitelon et. al 2016). For SOX10, we use RNA-seq data from a derivative of the S16 Schwann cell line in which CRISPR gene editing is used to eliminate the *Sox10* gene(Fogarty et. al 2020). In this knockdown data, the TF is downregulated to determine which genes had significant changes in their expression. We retrieve additional lists of knockdown data for SOX10 and EGR2 from (Srinivasan et al. 2012) and LOF data for SREBF1 from (Cermenati et al. 2015). We utilize various resources, such as Ensembl biomaRt(Durinck et al. 2009), g:Profiler(Reimand et al. 2007), and HGNC multi-symbol checker(Oh et al. 2022) to adjust gene names from mouse/rat to human names and retrieve updated gene names. Fold changes (FCs) < 0 imply that the TF may activate the TG, while FC > 0 may suggest that the TF represses the TG. For STAT1, we obtain data on comparative gene expression profiling analysis of RNA-seq data for differentiating STAT1 knockdown SCs and control SCs from NCBI GSE211336(Xu et al. 2023). We consider LOF genes for the TF as those where $|\log_2(\mathrm{FC})| > 0.5$ and the adjusted p-value < 0.05. Unfortunately, LOF data is missing for RXRG. LOF genes with nearby annotated ChIP-seq peaks for the TF are defined as validated direct TGs.

We validate novel SOX10 predicted TGs using SOX10 ChIP-seq (Lopez-Anido et. al 2015) epigenomic data of rats in the Peripheral Nervous System (PNS), including: SOX10 PNS Peaks, SOX10 Sciatic Nerve (SN) reads: Run 204, SOX10 PNS Filtered DBChIP Peaks (found a SOX10 site differentially expressed in SCs), Homer Filtered PNS K27ac, and H3K27Ac (marker for active enhancers (Zhang et al. 2020)) Immunoprecipitation Sorted Tag Directory; we have a visual session in UCSC genome browser available (Karolchik et al. 2009). Peaks files show called reads.

Cut&Run Data Processing: Cut&Run FastQ files are mapped to rat reference genome rn5 using Bowtie2 (Langmead et. al 2012; Langmead et al. 2009) to produce Binary Alignment Map (BAM) files. BAM files are filtered for mapped reads using BamTools(Quinlan and Hall 2010) and sorted into called peaks using MACS2(Zhang et al. 2008; Liu 2014). BedTools bamCoverage generates bedgraphs of Cut&Run and ChIP-seq samples. Bedgraph files show the distribution of reads. Heat maps are created via EAseq(Lerdrup et. al 2016). Data processing is performed in a cloud-based manner through GalaxyBiostars(Afgan et al. 2018). Cut&Run and ChIP-seq tracks are visualized using the University of California, Santa Cruz Genome Browser (Kent et al. 2002).

Co-binding / colocalization analysis among pairwise TFs: We perform co-binding analyses among all pairwise TFs for the 7 TFs above based on the rn5 rat ChIP-seq tracks for the TFs. For each $TF_i - TF_j$ TF-TF coordination pair, we determine the percentage of ChIP-seq peaks for TF_i that have binding overlaps with any given peaks for TF_j based on 15,864 H3K27ac peaks (histone epigenetic mark of active enhancer regions) shared between the PNS and S16 tracks. Similarly, the $TF_j - TF_i$ TF-TF coordination pair, determines the % of ChIP-seq peaks for TF_j that have binding overlaps with any given peaks for TF_i based on those shared H3K27ac peaks. We note that the value for $TF_i - TF_j$ may differ from that for $TF_j - TF_i$ based on the denominator.

In addition, we determine the Jaccard similarity (JS) among the TFs based on TF-TG regulatory network links for the TGs. We identify common TGs. Further, we overlay the regulatory regions in our TF-TG regulatory network (in hg38 human genomic coordinates) with the ChIP-seq peaks (after liftover from rat rn5 to hg38 coordinates) and determine TF-TG regulatory links for our core SC TFs that bind to regulatory regions with ChIP-seq support. We overlap the resulting TGs among TFs. In addition, we overlap the rn5 ChIP-seq peaks with the rn5 H3K27ac enhancers to determine enhancer-based ChIP-seq peaks and overlay those with our TF-TG regulatory networks to uncover how many shared predicted TGs have support based on ChIP-seq data for TFs in regions of active enhancers in rats. If there is a relatively large overlap in fraction of peaks between 2 different proteins from 2 respective ChIP-seq experiments, it may suggest that both proteins are cooperative in their gene regulatory behavior (Yu et al. 2015).

Ground truth gene regulatory network (GRN) in SCs: We note the following pooled results for SCs based on 17,049 TGs that are also found in the Dorsal Root Ganglion (DRG) SC gene expression data. This corresponds to 10,790 total unique TGs (including 4,708 unique TGs that are considered direct for at least 1 TF; nonetheless, they may be LOF TGs and/or direct TGs for other TFs). RXRG is missing from the analysis since we did not have any

LOF data for RXRG. This original GRN has 16,804 total TF-TG links, of which 6,393 links are TF-TG for validated direct TGs. In this original GRN, 75,523 links would have been possible among the 7 TFs and all LOF TGs (pooled LOF only TGs and valid direct TGs) and 32,950 links are possible among the 7 TFs and all validated direct TGs.

Ground Truth Gene Regulatory Network in Schwann Cells (Based on 17,049 Target Genes (TGs) found in Dorsal Root Ganglion (DRG) Schwann Cell Gene Expression Data): 16,804 total TF-TG links			
Transcription Factor (TF)	# LOF Only TGs	# Direct TGs	All LOF TGS **Total**
SOX10	5,404	2,278	7,682
YY1	2,838	1,273	4,111
STAT1	216	786	1,002
TEAD1	663	735	1,398
NR2F2	250	719	969
SREBF1	133	387	520
EGR2	907	215	1,122

Based on our input multi-omics analysis, we uncover results for 13,886 TGs in mSCs and the same 13,886 TGs in nmSCs. Now, instead of 16,804 total TF-TG links, we retail only 14,557 total TF-TG links in the ground truth GRN for both SCs, of which 5,740 links are TF-TG for validated direct TGs. This corresponds to 9,253 unique TGs (including 4,185 unique TGs that are considered direct for at least 1 TF; nonetheless, they may be LOF TGs and/or direct TGs for other TFs). Now, 64,765 links are possible among the 7 TFs and all LOF TGs (pooled LOF only TGs and validated direct TGs) and 29,290 links are possible among the 7 TFs and all validated direct TGs. **Results of TF-TG regulatory networks: mSCs:** 236149 rows originally in TF-TG regulatory network. Corresponding to 12,396 TGs and 221 TFs. After filtering: 183,242 rows now corresponding to 8,950 TGs and 221 TFs; **nmSCs:** 657,013 rows originally in TF-TG regulatory network. Corresponding to 13,454 TGs and 233 TFs. After filtering (rounding test MSE to 4 decimals): 277,541 rows now corresponding to 5,207 TGs and 228 TFs.

Final Ground Truth Gene Regulatory Network (GRN) in Schwann Cells (Based on 13,886 Target Genes (TGs) found in Dorsal Root Ganglion (DRG) Schwann Cell Gene Expression Data and Multi-omics Data-Derived Reference GRNs for SCs)			
Transcription Factor (TF)	# LOF Only TGs	# Direct TGs	All LOF TGS **Total**
SOX10	4,916	2,102	7,018
YY1	2,122	1,079	3,201
STAT1	198	713	911
TEAD1	526	652	1,178
NR2F2	228	660	888

SREBF1	102	327	429
EGR2	725	207	932

§B.1.10.6 Hypergeometric test of TF-TG regulatory network overlaps with known TGs for TFs in SCs

In mSCs and nmSCs, we perform hypergeometric testing of overlaps of predicted TGs with ground truth TGs to see if the predicted results are statistically significant. We do this for all LOF TGs and for the verified Direct TGs for the 7 SC TFs that we have experimental validation data for in rodents (e.g. rats). As a reminder, all LOF TGs comprise Direct TGs as well as LOF only TGs. Since we are missing corresponding LOF knockout data for RXRG, we cannot do this analysis for RXRG. We have 9,253 LOF TGs across 7 core SC TFs: EGR2, SOX10, NR2F2, STAT1, TEAD1, YY1, SREBF1. And, we have 7,018 LOF TGs for SOX10, 3,201 LOF TGs for YY1, 911 LOF TGs for STAT1, 1,178 LOF TFs for TEAD1, 888 LOF TGs for NR2F2, 429 LOF TGs for SREBF1, and 932 LOF TGs for EGR2. We also find 4,185 validated Direct TGs across these 7 core SC TFs, where we have: 2,102 for SOX10, 1,079 for YY1, 713 for STAT1, 652 for TEAD1, 660 for NR2F2, 327 for SREBF1, 207 for EGR2. We run NetREm for mSCs and nmSCs and find how many total TGs we have for each of these 7 core SC TFs. Then, we overlap these total TGs with LOF TGs for the TF for the LOF analysis (and overlap total TGs with validated Direct TGs for the TF for validated direct TG analysis). For instance, for SOX10 LOF TGs in mSCs, we use the hypergeometric test to determine whether the overlap between the SOX10 predicted mSC TGs and the SOX10 LOF TGs is greater than what would be expected by random chance.

- Null Hypothesis of Hypergeometric Test is: H_O: there is NO statistically significant association between the predicted SOX10 TGs and the LOF TGs for SOX10. Any overlap observed is due to random chance.
- Alternative Hypothesis is: H_A: there is A statistically significant association between the predicted SOX10 TGs and the LOF TGs for SOX10. Any overlap observed is not due to random chance.

If there are k common TGs between SOX10 mSC TGs and all LOF TGs for SOX10. This is the overlap. The p-value from the hypergeometric test represents the probability of observing at least k common TGs (or a more extreme overlap) between the predicted SOX10 TGs in mSCs and the SOX10 LOF TGs, assuming that the Null Hypothesis is true. We use the p-value to determine if overlaps are statistically significant. This helps us determine if the predicted TGs in SOX10 in mSCs are over-enriched (for NetREm's TF-TG regulatory networks in mSCs).

For the hypergeometric test, there is P, the total population (total # of items in the population, which here is the total # of unique LOF TGs among the 7 SC TFs: 9,253 TGs). So $P = 9{,}253$. We also see the # of SOX10 LOF TGs in the ground truth network, which is $m = 7{,}018$. We look at the # of SOX10 mSC TGs that are also SOX10

LOF TGs (i.e. the overlap k), which in our case is $k = 564$ TGs. We also look at the LOF TGs that are found as core for any SC TF, which is the sample size n: 709 TGs. We then run the hypergeometric test with # of successes $k = 564$ TGs, sample size $n = 709$ TGs, # of successes in the population $m = 7{,}018$ TGs, population size $P = 9{,}253$ TGs. Statistically significant results will show that our set of SOX10 TGs in mSCs is overenriched SOX10 LOF TGs, which will be a good sign. In fact, we get "over enriched 1.05 fold compared to expectations", with a p-value of 0.00855 that is statistically significant over-enrichment. We do this for all 7 core SC TFs in mSCs and in nmSCs, for LOF TGs and for validated Direct TGs.

§B.1.10.7 Contextual PPI Database (CPPID)

We employ an integrated CPPID (Kotlyar et. al 2022) of dummy (0 = absent, 1 = present) values, annotated for various diseases/tissues/traits, as a validation tool for predicted TF-TF links. We assigned 0 to any links that are missing from the contextual PPI; thus, links with 0 value may be a True Negative (TN) or currently unknown. This is another example of positive unlabeled learning; to adjust for this, we undersample majority class 0. For machine learning analysis tasks, we use B scores of TF-TF links for each TG in a matrix. We have a matrix of TF-TF link coordination scores (i.e. embeddings) as rows, for respective TG as columns. Overall, NetREm provides a holistic approach to understanding cell-type-specific GRN mechanisms and enriches our understanding of core cell-type interactions (direct/indirect) among TFs involved in these processes.

SCs: We focus on CPPID links annotated for Peripheral Nervous System (PNS) neoplasms, and other nervous system-related diseases or disorders. *AD:* We use this CPPID, which categorizes these links as either associated with AD or currently unassociated. In addition, we employ machine learning to identify TGs predictive of neurodegenerative diseases across eight cell types. For this task, we opt to focus on neurodegenerative diseases at large, rather than AD specifically, to expand the pool of positive links.

Next, for each cell-type, we obtain the B matrix of coordination scores representing TF-TF links for each TG in the AD stage: $\widetilde{B_{AD}}$. We also obtain $\widetilde{B_{Control}}$, which refers to the B matrix of TF-TF coordination scores for each TG in the Control stage for that cell-type. We note that values in $\widetilde{B_{AD}}$ and $\widetilde{B_{Control}}$ will be between -100 and 100 based on our definition of coordination scores. Thus, we do not perform any scaling on these coordination scores and note that both matrices should be on a comparable scale of values. Then, we perform 0 padding to ensure that both $\widetilde{B_{AD}}$ and $\widetilde{B_{Control}}$ matrices have the same dimensions corresponding to the same TF-TF links and TGs, where TF-TF links are rows and TGs are columns. We then compute $\widetilde{B_{Final}} = \widetilde{B_{AD}} - \widetilde{B_{Control}}$, to obtain our final

matrix of net changes in TF-TF coordination scores for the cell-type TGs. That is, these differences in B values between AD and controls across TGs are used to create $\widetilde{B_{Final}}$. Using the CPPID, we label these TF-TF links as either related to neurodegenerative diseases (class 1) or not (class 0). Due to class imbalance and potential for false negatives (FNs) in class 0, we undersample majority class 0 to create a balanced dataset for each cell type (via imblearn.undersampling RandomUnderSampler (Lemaître et al. 2017)). We train various machine learning models based on this balanced data.

We train a Random Forest Clsasifier model using all defaults (e.g. 100 trees, gini estimator for split). We also train XGBoost classifier models (via sklearn XGBClassifier (Pedregosa et al. 2011)) with the logloss evaluation metric and all other defaults. Further, we train Naïve Bayes, SVM, and Logistic Regression models using the respective defaults. Model performance is evaluated using stratified 5-fold Cross Validation (CV), with metrics for each fold, including: accuracy, Receiver Operator Characteristic (ROC) curves, and Precision-Recall curves. We report averaged values across all 5 folds for: area under the ROC curve (AUC), area under Precision-Recall curve (AUPR), and accuracy. Our final classifiers are trained on the entire dataset for each cell type, and we filter the genes to select the top 500 most influential TGs for each cell type (based on highest feature importance gain values), which may serve as potential biomarkers for neurodegenerative diseases, warranting further investigation. We conduct Gene Enrichment Analysis to ensure these identified TGs are biologically meaningful and relevant.

Further, we use the CPPID to determine any biologically-meaningful annotations for top cell-type TF-TF coordination links $\bar{B}$. Currently, this CPPID does not have cell-type PPI annotations. Nonetheless, our findings may show NetREm may typically prioritize cell-type and/or context-specific TF-TF coordination links. That is, this is a proxy for determining how cell-type-specific our TF-TF coordination networks can be (i.e. potential cell-type TF-TF PPINs). For the given cell-type, we retrieve cell-type TF-TF links that are relatively strong $|\bar{B}| \geq 85$. We overlay those with annotated CPPID links, computing the % of top TF-TF links enriched for each of the 243 different annotations. We rank the results, where low ranks represent a greater %.

§B.1.10.8 Gold Standard Gene Regulatory Networks (GRNs)

We have gold standard validation data for human embryonic stem cells (hESCs), Human Hematopoietic stem Cells (HSCs), mouse embryonic stem cells (mESCs). In hESCs, we have the 5,050 TF-TG signed ground truth GRN links subset from the original atlas of regulatory links(Sharov et al. 2022) for the 1,250 randomly-selected TGs. In addition, we have the sign (+: activate, -: repress), which we can use to evaluate the c^* sign (+ or -) for predicted TF-

TG regulatory links. We can use this ground truth GRN for validation as it is the underlying input data for SERGIO to generate gene expression data. In HSCs and in mESCs, we utilize the pooled gold standard GRN networks from (Zhang et al. 2023). These are the following gold standard networks that we pool for HSCs: UniBind.gs_network, Cus_ChIP_2, Cus_KO_0.01_Cus_ChIP_2_intersect, Cus_KO_0.01, Cus_KO_0.01_union_intersect. For mESCs, we gather the gold standard networks from mESC_KDUnion, mESC_chipunion, and mESC_chipunion_KDUnion_intersect that (Zhang et al. 2023) mined and curated from experimentally derived networks of regulatory interactions from literature and from existing scientific databases.

Total Potential TF-TG regulatory links that can be predicted:

All TF-TG regulatory network predictions can be classified into these 4 categories, which form a confusion matrix.

- *True Positives (TP):* the # of instances where the model correctly predicts the positive class. In other words, it is the # of correct predictions that an instance is positive.
- *False Positives (FP):* the # of instances where the model incorrectly predicts the positive class. That is, it is the # of times the model predicts an instance as positive when it is actually negative.
- *False Negatives (FN):* the # of instances where the model incorrectly predicts the negative class. It means the # of times the model predicts an instance as negative when it is actually positive.
- *True Negatives (TN):* the # of instances where the model correctly predicts the negative class. It is the # of correct predictions that an instance is negative.

TF-TG regulatory links (and/or TF-sign-TG regulatory links) that are not predicted by a model are predicted negatives (as they are missing links in the model's inferred TF-TG regulatory network). These missing links may be FNs (if they are true links in the ground truth network that should have been predicted) or TNs (not found in the ground truth network and correctly not predicted as a meaningful link by the model). Those regulatory links that are predicted by a model are predicted positive (and may comprise TPs if they are correct links found in the ground truth network or FPs if they are predicted links that are actually not found in the ground truth network). To calculate the total number of TF-TG regulatory links that are theoretically possible for each scenario, we utilize the following formula: Total # of predictions = Total TF-TG regulatory links possible = TPs + FPs + TNs + FNs = $\big((Total\ \#\ of\ TFs) \times (Total\ \#\ of\ TGs)\big) - (\#\ of\ TFs\ that\ are\ TGs)$. Then, for instance, in hESCs (1,250 TGs and 207 TFs, of which 15 TFs are also TGs), we have: (207)(1,250) – 15 = 258,735 potential TF-TG regulatory links that can be predicted. Thus, TP + FP + FN + TN = 258,735. Likewise, in mESCs (19,225 TGs and 195 TFs, of

which 194 TFs are also TGs), we have 3,748,681 potential TF-TG regulatory links that can be predicted. Further, in mDCs (9,087 TGs and 93 TFs, where all TFs are also TGs), we have 844,998 potential TF-TG regulatory links that can be predicted. TFs that are not TGs in gene expression data sets may be considered potential master regulators.

Metrics used when ground truth Gene Regulatory Networks (GRNs) are available:

We evaluate the TF-TG regulatory network predictions inferred by various models: NetREm (with hyperparameters β and α altered as needed), ElasticNetCV, LassoCV, Linear Regression, RidgeCV, and GRNBoost2 (Pedregosa et al. 2011; Moerman et al. 2019). For hESCs, we have information on the sign of the TF-TG regulatory link; that is, we have positive (+) signs for activator TFs and negative (-) signs for repressor TFs. In hESCs, we refer to True Positive (TP) predictions (correct TF-TG regulatory links predicted) as TF-TG regulatory links that are not only found in the ground truth validation atlas but also have the correct coefficient $c^*_{TF_i}$ where $c^*_{TF_i} > 0$ if TF_i activates TG and $c^*_{TF_i} < 0$ if TF_i represses TG. That is, in hESCs, we can also focus on TF-sign-TG regulatory links where sign can be (+ or -). For the TF-sign-TG regulatory link comparison in hESCs, we cannot consider GRNBoost2 as that only returns relative importance scores for TF-TG scores and does not provide any signed coefficient values; instead, we consider NetREm the 4 benchmark models (ElasticNetCV, LassoCV, Linear Regression, RidgeCV).

For HSCs and mESCs, we have TF-TG regulatory links but lack information regarding the potential role of the TF in terms of the TG expression. Thus, our TP predictions are simply predicted TF-TG regulatory links that are found in the gold standard GRN networks. We also analyze TPs in hESCs in terms of TF-TG regulatory links, ignoring the coefficient c^* signs. When assessing the TF-TG regulatory links, we compare NetREm with the 5 benchmark models (including GRNBoost2). There are various metrics we can use to evaluate the model predictions of TF-sign-TG regulatory links (in hESCs) and TF-TG regulatory links (in hESCs, HSCs, mDCs, and mESCs).

- Precision measures the proportion of predicted positives (True Positives (TPs) + False Positives (FPs)) that are correct (i.e. TPs) (out of all of the regulatory links, how many have been validated). Precision = $\frac{TP}{TP + FP} = \frac{TP}{Total\ Predicted\ Positives}$ = True Positive Rate (TPR)
- Specificity measures the proportion of actual negatives (True Negatives (TNs) + False Positives (FPs)) that are correctly identified. Specificity = $\frac{TN}{FP + TN} = \frac{TN}{Total\ Actual\ Negatives}$

- Recall measures the proportion of actual positives (TPs + FNs) that are correctly identified. That is, it is the # of correct positive predictions (i.e. TPs) divided by all the relevant samples (i.e. all samples that are truly positive).

  Recall = $\frac{TP}{TP + FN} = \frac{TP}{Total\ Actual\ Positives}$ = Sensitivity

- F1 Score is the harmonic mean of precision and recall. $F1 = (2)\left(\frac{(\text{precision}) \times (\text{recall})}{(\text{precision}) + (\text{recall})}\right)$

- Balanced Accuracy (BACC) is great for skewed datasets. $BACC = \frac{(\text{sensitivity}) + (\text{specificity})}{2}$

- Total correct predictions = TPs + TNs. Total errors = FPs + FNs

- Overall accuracy is the ratio of total correct predictions over the total number of predictions. And the total # of predictions is the total # of possible regulatory links. $= \frac{TP+TN}{TP+TN+FP+FN}$

§B.1.10.9 RTNduals Benchmark analysis of TF-TF coordination scores:

We evaluate and compare the potential reliability and accuracy of predicted TF-TF interactions by NetREm in the context of other tools, such as RTNduals (Chagas et al. 2019). Like NetREm, RTNduals reconstructs Transcription Networks (via its parent package RTN (Groeneveld et al. 2023)) and assesses whether regulons (i.e. TFs) might cooperate or compete within the regulatory frameworks to form possible co-regulatory loops (i.e. forming dual regulons). RTNduals analyzes co-regulatory associations (i.e. loops) among regulons, uses the Data Processing Inequality (DPI) algorithm (Meyer et. al 2008) to remove the weakest interaction between 2 regulators and a common TG, and ultimately RTNduals infers statistically significant dual regulons from input gene expression data and a list of potential TFs. Both tools (RTNduals and NetREm) recognize that each TG in a given TF-TG regulatory network may be linked to multiple TF regulators based on direct and/or indirect interactions among TFs.

1. Comparative Analysis Approach

We discern whether integrating a PPI network as a prior knowledge network constraint enhances NetREm's predictive accuracy for TF-TF links (direct and/or indirect) over RTNduals, using identical input gene expression data and TFs list. We assume that this PPI ground truth network is static across various cell types and conditions, eliminating the need for users to curate additional PPI data, unless desired. For a balanced comparison, both RTNduals and NetREm are run on the same input training gene expression data and a fixed list of TFs. We then compare their predicted TF-TF links, focusing on the same top number of interactions.

2. Utilizing Empirical Data

To fortify our analysis, we use our updated comprehensive Protein-Protein Interaction Network (*NewNet*) in humans (4,760,982 unique protein-protein edges based on gene names), which is based on STRINGdb version 12 and other data resources that we detail in **§B.8**. Please note that we use NewNet as our PPIN (from which we subset our TF-TF PPIN) for our Schwann cell and AD applications. We also add some newer TF-TF links (missing form this PPIN) based on a recent study(Göös et al. 2022) investigating human TF protein interaction networks via rigorous physical experimental analysis of 109 different proteins (1,431 TFs), to yield our *EvaluateNet*. This study recognizes interactions with a SAINT (Significance Analysis of INTeractome) score above 0.75, as assessed by SAINTexpress software version 3.1.0[108], as being significant and highly-confident. It uncovers 6,503 and 1,336 strong direct TF-TF interactions through BioID and AP-MS analysis respectively, with 200 common interactions, many of which are novel and absent in input PPI. That is, in (Göös et al. 2022) returns TF-TF direct interactions with high-confidence based on BioID and/or AP-MS experiments with strong SAINT scores: Highly Confident Novel Human TF-TF Protein Interaction Networks based on BioID and/or AP-MS physical experiments. We also add indirect/direct PPI links from BioGrid 2024 release that are not in NewNet. These links found in (Göös et al. 2022) and (Oughtred et al. 2021) that are not in our comprehensive PPIN (*NewNet*) will be the novel ones for *NewNet* evaluation. That is, *EvaluateNet* simply annotates the TF-TF links that are found in *NewNet*, (Göös et al. 2022) and (Oughtred et al. 2021) as annotated TF-TF links. Thus, annotated links could be: NewNet links (indirect and/or direct links), (Göös et al. 2022) links only (verified direct, physical TF-TF interactions by BioID and/or AP-MS experiments), BioGrid new links (Oughtred et al. 2021) including direct/indirect interactions. We will use the Version 11 STRINGdb PPIN (updated from January 2019 to October 2020) to train NetREm (i.e. NetREm v11). We use *NewNet* to train NetREm (i.e. NetREm NewNet). Thus, we will have the following buckets for TF-TF coordination scores:

- Known Protein-Protein Interactions (PPIs)
    - In Both STRINGdb V11 and *NewNet*
    - Novel Discoveries: not in V11 network but in NewNet
- Poor results:
    - Still unknown (not in V11 networks and not in *EvaluateNet*) → may be True Negatives (TNs) or False Negatives (FNs).
    - Removed (in V11 networks but are not in *EvaluateNet*) → potential False Positives (FPs)

3. <u>Methodological Approach</u>

For each cell type and/or condition, we scrutinize our existing GRN knowledge for all respective TGs, identifying the set of final potential cell-type candidate TFs $\mathcal{N}$ present in our gene expression data. These would be potential cell-type TFs that we derive from multi-omics data. These are then overlapped with the 1,431 TFs to form the input set of TFs. These TFs are provided into RTNduals and NetREm. Next, we provide these TFs to RTNduals (with defaults: 5% p-adj, 1,000 permutations, 100 bootstraps, epsilon eps = NA: estimate the threshold from the empirical null distribution computed in permutation and bootstrap steps to compute the DPI-filtered regulatory network).

We run NetREm for the TGs in the cell-type for the fixed # of candidate TFs $N = \mathcal{N}$; if the TG is also a TF, we just remove that TF from the list of candidate TFs for that TG, so that TG has $N = \mathcal{N} - 1$ candidate TFs (**Table B.16**). That is, we do not use any other prior GRN knowledge. We fit NetREm with $\beta = 1$, no y-intercept, and use LassoCV to optimize α using 5-fold Cross Validation (CV) to minimize the CV Mean Square Error. In any case, we note that NetREm may incorporate TG-specific knowledge, which RTNduals may be unable to do. Nonetheless, we want to strictly compare the performance of RTNduals with that of NetREm in the absence of additional TG-specific knowledge. Thus, we use the same TFs for each TG for the sake of comparison with RTNduals. We also note a limitation of RTNduals is that it may not return any significant results for some of the single-cell expression data; unlike NetREm, RTNduals is not adept at single-cell data, unlike NetREm that can handle both bulk and single-cell expression (sparse!) data. Thus, we will focus on the situations where RTNduals returns predicted TF-TF links. For each TG, we learn the TG-specific TF-TF coordination network B. Then, we aggregate and average the results across all TGs to learn our final $\bar{B}$ matrix of cell-type coordination scores. which also represents the average B matrix for all TGs in this application (given that we are using the same TFs for each TG for the sake of comparison with RTNduals). We note our $\bar{B}$ matrix always returns more predicted TF-TF links than RTNduals does, which could be up to: $(\mathcal{N})(\mathcal{N}-1)(0.5)$ for NetREm. We run NetREm using the older *V11* STRINGdb PPIN (i.e. V11). We also run it using the *NewNet* of updated knowledge (V12 and beyond!).

- NetREm($\beta = 1$, LassoCV determines α, input PPIN: older V11 network).
- NetREm($\beta = 1$, LassoCV determines α, input PPIN: *NewNet*).

Our goal is to determine if RTNduals can pick up on true TF-TF relations with PPI support and compare results against NetREm, which incorporates either an outdated PPIN (with some false information) or a more updated network (*NewNet*).

4. Evaluation and Benchmarking

We assess the number k and quality of significant TF-TF links predicted by RTNduals and subsequently align our $|\bar{B}|$ matrix (magnitude of cell-type TF-TF coordination scores, where higher values in $|\bar{B}|$ denote stronger antagonism and/or cooperativity among TFs). Thus, we include the top k TF-TF links (based on TF-TF coordination scores for $|\bar{B}|$) for a harmonized comparison. The total number and respective percentages of TF-TF links, annotated based on *EvaluateNet*, are compared between RTNduals and NetREm (using *V11*) and NetREm (using *NewNet*). This comparison enables us to critically assess and benchmark the efficacy of NetREm's predicted TF-TF coordination network against RTNduals, based on identical input data. This comparative analysis is poised to contribute valuable insights into the reliability of the top TF-TF coordination links predicted by NetREm.

§B.1.10.10 Neural Cell TF-clusters ChIP-seq tracks

We use raw data for the ENCODE3 TF clusters track (**encRegTfbsClusteredWithCells.hg38.bed.gz**) from UCSC Genome Browser (Transcription Factor ChIP-seq Data Standards and Processing Pipeline – ENCODE (encodeproject.org)). We analyze our results for cell-type TF-TF cooperativity in Control Excitatory Neurons and Control Inhibitory Neurons (from Application 7) for these 6 TFs with available data in neural cells (CTCF: EP300, EZH2, MXI1, RAD21, SMC3). We determine Jaccard Similarity of overlaps between each pair of TFs. We consider even 1 base pair overlap as important and perform overlap analysis using GenomicRanges (Lawrence et. al 2013).

$$JS(TF_i, TF_j) = \frac{\#\text{ of ChIP} - \text{seq peaks with overlap in binding for } TF_i \text{ and } TF_j}{\#\text{ of peaks bound by } TF_i \frac{\text{and}}{\text{or}} TF_j}$$

2 ChIP-seq experiments by 2 different TF proteins can be analyzed to see if they have a large fraction of peaks with overlap; if so, these TFs may be cooperative in terms of their regulation (Yu et al. 2015).

§B.1.10.11 Validate TF-TF Coordination Prediction Scores for True Signals using new PPINs

We evaluated TF-TF coordination prediction scores using new PPINs in 4 real-world applications with available gold standard TF-TG validation data: mouse embryonic stem cells (mESCs), simulated human embryonic stem cells (hESCs), human hematopoietic stem cells (HSCs), mouse dendritic cells (mDCs).To benchmark the performance of our predicted cell-type TF-TF coordination scores $\bar{B}$, we used 2 versions of the STRING database: an older version (v11) as a proxy for historical data and a newer version (v12) as current information to gauge NetREm's ability to predict future links. The evolution of PPINs from v11 to v12 provides a basis for validation. For example, in mESCs, our fully connected TF-TF PPIN includes both known (v11: 5,044; v12: 6,856) and artificial (v11: 13,871;

v12: 12,059) TF-TF edges for 195 candidate TFs. Additionally, 25 singleton TFs in v11 have known PPIs in v12. We classified discoveries as follows:

- **Valid discoveries (False Negatives, FNs):** Found only in v12, demonstrating NetREm's ability to predict future PPI links.
- **Unknown links:** Absent in both v11 and v12; these could be future discoveries (FNs) or true negatives (TNs).
- **V11 links:** Present in v11; they may persist in v12 (Known) or be removed (False Positives, FPs).
- **Non-V11 links:** Absent in v11; these are artificially added with default edge weight and could be FNs (discovered in v12 or beyond) or TNs (truly unknown).

We compare the magnitude of cell-type TF-TF coordination scores ($|\bar{B}|$) across these groups using one-sided Welch statistical tests (see **Table B.7**) to identify statistically significant differences. The following comparisons are made:

1. Known TF-TF links (v11) vs. Unknown links (v11).
2. Known and Retained TF-TF links vs. Known and Removed TF-TF links.
3. Valid discovery TF-TF links (not in v11 but in v12) vs. Unknown links (not in v11 and v12).
4. Valid discovery TF-TF links (not in v11 but in v12) vs. Removed links (in v11 but not in v12).
5. Known TF-TF links (v11) vs. Valid discoveries (not in v11 but in v12).
6. Known and Retained TF-TF links (v11) vs. Valid discoveries (not in v11 but in v12).
7. Unknown TF-TF links (not in v11 or v12) vs. Removed links (in v11 but not in v12).
8. Unknown TF-TF links (not in v11 or v12) vs. Removed links (in v11 but not in v12) (reverse comparison).

Key Findings: Our findings underscore NetREm's capacity to discern between different types of TF-TF links, demonstrating a nuanced understanding of PPI network evolution:

- *Prioritization of True Positives:* $\bar{B}$ is significantly higher for TF-TF links identified as novel discoveries in v12 compared to those removed (indicative of FPs) in the transition from v11 to v12. This suggests NetREm's effectiveness in predicting meaningful future TF-TF interactions.
- *Recognition of Known Links:* Across multiple scenarios, including mESCs and (mDCs with varying sparsity parameters α, β fixed at 1), $\bar{B}$ consistently favored known links over unknown or newly discovered ones, highlighting NetREm's reliance on established knowledge.

- *Future Discovery Potential:* Despite its preference for known interactions, NetREm also demonstrates a significant capability to predict valid, yet previously unknown TF-TF links, as evidenced by higher $\bar{B}$ scores for discoveries in v12 compared to persistently unknown links. Thus, there is some transfer learning that takes place.

Our detailed exploration and benchmarking of NetREm against evolving PPI databases reveal its sophisticated ability to navigate and predict the complex landscape of TF-TF interactions, affirming its value as a powerful tool for understanding cellular mechanisms and their evolution over time.

Section §B.1.11: Software Implementation of NetREm

Please note implement NetREm on GitHub as an open-source software package in Python:  https://github.com/SaniyaKhullar/NetREm/tree/main.

§B.1.11.1: Handling input gene expression data

By default, NetREm uses 70% of the gene expression data for training the model and the remaining 30% for testing. Other default settings for NetREm include standardizing the X variable and standardizing the y variable. NetREm employs StandardScaler from the sklearn.preprocessing package(Pedregosa et al. 2011) to standardize X by fitting the StandardScaler (i.e. computing the mean and standard deviation of each feature) on X_{train}. Then, these calculated parameters are used to perform the transformation, i.e., subtracting the mean and dividing by the standard deviation, on both X_{train} and X_{test} sets. Similarly, to standardize y, NetRem fits the StandardScaler (i.e. computing the mean and standard deviation of each feature) on y_{train}. Then, these calculated parameters are used to perform the transformation, i.e., subtracting the mean and dividing by the standard deviation, on both y_{train} and y_{test} sets. These operations help convert $X^{(0)}$ and $y^{(0)}$, which are typically gene expression data and in measurements like FPKM, to unitless quantities X and y, respectively.

Standardizing a variable involves subtracting the mean and dividing by the standard deviation (which is positive number). For these M cell samples (rows), our $X^{(0)} \in \mathbb{R}^{M \times N}$ matrix contains single-cell gene expression data for these N TFs (columns), while $y^{(0)} \in \mathbb{R}^{M}$ is the expression vector for our given TG. Again, the dimensions of $X^{(0)}$ may vary across TGs as the number of candidate TFs, N, may differ for TGs if we incorporate prior knowledge on cell-type gene regulation for initial feature selection. We note that this single-cell gene expression data may have been subject to batch correction and other pre-processing using toolkits like Seurat (e.g. normalizing

molecular counts to consider impact of variance in sequencing depth across cells that may be of different cell-types. For instance, Seurat's LogNormalize method divides counts for each gene in a cell by the total counts for that cell (multiplied by a scaling factor, e.g. 10,000) to yield a value; then $\log_e(\text{value} + 1)$ is calculated for each gene to yield the normalized counts (e.g. log-transformed counts per 10,000 total counts). In addition, Seurat's scaleData scales each gene to have a mean of 0 and variance of 1 across cells; nonetheless, we do not make assumptions regarding prior pre-processing or units of our original gene expression data $X^{(0)}$ and $y^{(0)}$.

When we run NetREm, we randomly partition our M samples into mutually exclusive training and testing samples and select the corresponding $(X^{(0)}, y^{(0)})$ gene expression data for our respective training $(X^{(0)}_{train}, y^{(0)}_{train})$ and testing data $(X^{(0)}_{test}, y^{(0)}_{test})$. We will standardize our datasets utilizing the training data as a guide to ensure robust model generalization to unseen data by preventing data leakage from the test set during the preprocessing stage. The user can opt to center the response y variable ($y - \bar{y}$, where $\bar{y}$ is the mean of y), which is False by default. If the user opts to center the y variable, then NetREm will compute $\overline{y_{train}}$ and subtract that from the response variables in both y_{train} and y_{test} sets to center y. This implementation thus utilizes the training data as a guide to ensure robust model generalization to unseen data by preventing data leakage from the test set during the preprocessing stage.

§B.1.11.2: Running *netrem* and *netremCV* functions

NetREm utilizes 2 hyperparameters: β for network constraint and α for sparsity. These can either be optimized dataset-specifically through cross-validation (CV) or manually set by the user. The chosen values strike a balance between the constraints imposed by the hyperparameters and the fit to the data. Lasso regression is performed on $\tilde{X}$ and $\tilde{y}$ using Python's scikit-learn package (Pedregosa et al. 2011). We provide 2 distinct functions: *netrem* and *netremCV*, each with its own strategy for hyperparameter selection. Generally, the choice of β influences the optimal value for α, given that α depends on the embeddings $\tilde{X}$ and $\tilde{y}$, which are scaled by β.

That is β is the network-constrained hyperparameter and it defines network regression gene embeddings $\tilde{X}$ and $\tilde{y}$. It does this by impacting the weight of the prior network knowledge (i.e. A that is based on input PPI network (PPIN)) in determining $E = \frac{X^T X}{M} + \beta A \in \mathbb{R}^{N \times N}$, upon which Singular Value Decomposition (SVD) is performed to then help us eventually obtain our network regression gene embeddings: $\tilde{X}$ and $\tilde{y}$. Thus, we often select β first and then may use LassoCV solvers to optimize α based on $\tilde{X}$ and $\tilde{y}$. Traditional Lasso solvers in sci-kit learn use α to denote the sparsity prior for Lasso $\mathcal{L}_1$ regularization. NetREm inherits from sklearn.base

RegressorMixin and *BaseEstimator* classes and is a Sci-kit Learn estimator itself. Hence, for consistency and to avoid confusion, we retain α for our sparsity prior. Nonetheless, β is the 1st hyperparameter selected and α is conditioned to operate based on embeddings generated for the given β. Hence, contrary to the Greek alphabet, our order is sequential for our hyperparameters: $\beta \rightarrow \alpha$. We ensure the list of N candidate TFs for the TG does not contain TG as a TF (that is, we ensure there are no overlaps in variables in X and y).

Function: *netrem*

The *netrem* function offers multiple approaches for determining β and α and is more flexible. Users have the option of employing GridSearchCV (Pedregosa et al. 2011) or Bayesian optimization via the Scikit-Optimize (skopt) library with Gaussian process minimization Alternatively, users can manually input β (or use the default) and decide on α. If α is either provided or set to default, NetREm employs Lasso for model fitting; else, NetREm uses LassoCV to automatically optimize α via CV.

Function: *netremCV*

Contrastingly, *netremCV* employs CV to automatically optimize both β and α, although at a higher computational cost. Nonetheless, there are cases in which *netrem* may outperform *netremCV* or vice-versa; please assume that unless otherwise mentioned, *netrem* is used for the applications in this paper.

§B.1.11.3: ElasticNetREm implementation: *elasticnetrem* function

We have an ElasticNetREm implementation of NetREm, which changes the objective function based on $\tilde{X}$ and $\tilde{y}$ to be solved via ElasticNet or ElasticNetCV instead of Lasso or LassoCV. In addition, there are other transformations that may utilize the updated $\tilde{X}$ and $\tilde{y}$ latent gene embedding representations, in a manner that combines the penalties of Lasso ($\mathcal{L}_1$norm) and Ridge ($\mathcal{L}_2$ norm) regularizations. NetREm may be also solved via ElasticNet regression, with the following equation that is adapted from $Eq(2)$.

$$c^* = \underset{c}{\operatorname{argmin}} \tilde{f}(c) = \frac{1}{2N}\left|\left|\tilde{y} - \tilde{X}c\right|\right|^2 + \alpha\rho\left|\left|c\right|\right|_1 + \frac{\alpha(1-\rho)}{2}\left|\left|c\right|\right|_2^2$$

This originates from the corresponding network regularization optimization formula: ElasticNetREm:

$$c^* = \underset{c}{\operatorname{argmin}} f(c) = \frac{1}{2M}\left|\left|y - Xc\right|\right|^2 + \alpha\rho\left|\left|c\right|\right|_1 + \frac{\alpha(1-\rho)}{2}\left|\left|c\right|\right|_2^2 + \frac{\beta}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(\frac{c_i}{\sqrt{d_i}} - \frac{c_j}{\sqrt{d_j}}\right)^2$$

Here ρ is the parameter that controls the balance between $\mathcal{L}_1$ and $\mathcal{L}_2$ penalties (i.e. $\mathcal{L}_1$ ratio), α is the overall strength of the regularization, $\left|\left|c\right|\right|_1$ is the $\mathcal{L}_1$ norm (i.e. $\sum_{i=1}^{N}|c_i|$), and $\left|\left|c\right|\right|_2^2$ is the $\mathcal{L}_2$ norm squared of the

coefficients vector c (i.e. $\sum_{i=1}^{N} c_i^2$). We note that $0 \leq \rho \leq 1$, where $\rho = 1$ for Lasso ($\mathcal{L}_1$ penalty only) and $\rho = 0$ for Ridge ($\mathcal{L}_2$ penalty only).

§B.1.11.4: Handling directed input prior networks

NetREm is most effective when applied using positive-weighted, undirected prior networks as constraints. To assist users in converting directed networks into undirected similarity networks among nodes, we implement a weighted Node2Vec network embedding algorithm (Cohen 2023; Grover and Leskovec 2016). This algorithm conducts random walks in the directed network, inputs these walks into a skipgram model (neural-network that predicts the likelihood of the content nodes given the current nodes) to learn node embeddings, and then calculates the cosine similarity among the nodes based on the embeddings.

§B.1.11.5: Running NetREm for a cell-type (many target genes)

Simprem function (no prior GRN knowledge)

In the absence of prior gene regulatory network (GRN) knowledge, we will have a fixed set and # of candidate TFs for each TG: $N = \mathcal{N}$, where $\mathcal{N}$ is the set of potential cell-type TFs. In the cases where the TG is also one of the TFs in $\mathcal{N}$, we will just remove that TF from the list of TFs for that TG, so that TG will have $N = \mathcal{N} - 1$ candidate TFs. When we run NetREm for each TG (step-by-step in an iterative loop), we can optimize its performance by realizing that the underlying input TF-TF PPIN network would be the same for a majority of cases (where TGs are not TFs), so we can reuse that network each time (instead of spending time preprocessing and recreating the same TF-TF PPIN network (W) and deriving V, D, A) and only make changes to the TF-TF PPIN network for those cases where the TG is also a TF. Thus, we build out the simprem function to keep things simpler. The *simprem* function helps us run NetREm faster in the absence of prior knowledge.

netremany function (with prior GRN knowledge)

This is a portmanteau of "NetREm" and "many" and is a capsule function that allows us to run NetREm when there may be prior GRN knowledge for some TGs, where we have some tailored candidate TFs for them. It enables the user to run NetREm for dozens or thousands of TGs in the cell-type in 1 function. Naturally, it will be slower than *simprem* because it is recreating the TF-TF PPIN for each TG in the cell-type, which takes time.

§ B.2 Supplementary figures

Figure B.1 We elaborate on **Fig. 3.2**, and provide our framework for analyzing NetREm outputs when building our TF-TG regulatory networks (complementary GRNs) and TF-TF interaction networks.

While the demonstration example illustrates how NetREm can be applied to solve a range of problems in various fields, we focus on applying NetREm in biology (to uncover cell-type-specific interactions among Transcription Factors (TFs) that are linked to the regulation of respective target genes (TGs)). We have 5 predictors (TFs, X variable: respective expression levels) and response variable (TG with expression y).

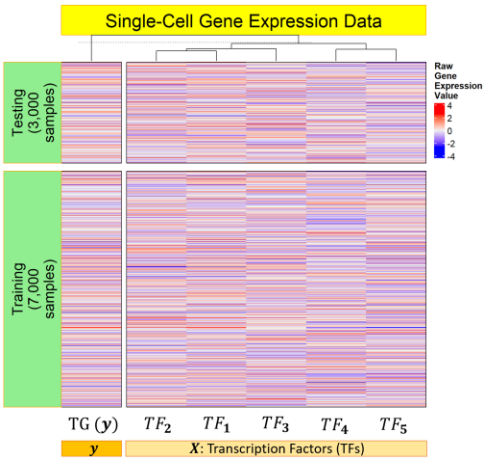


Figure B.1A) - Heatmap of the single-cell gene expression data of 10,000 cell samples prior to standardizing it based on training data.

This data has ~ 40% sparsity (i.e. 0 values) for each of the 6 variables ($TF_1, TF_2, TF_3, TF_4, TF_5, y$) to resemble the current realities of single-cell data more closely. The 5 predictors in X are clustered hierarchically with the dendrogram provided above the plot, such that TFs 1 and 2 cluster together (then with TF_3) and TFs 4 and 5 cluster together based on their respective expression profiles and expression dynamics. The # of cell samples used to train our models is 70% of 10,000 so M = 7,000. Since $N = 5$, this is a scenario where $N \ll M$.

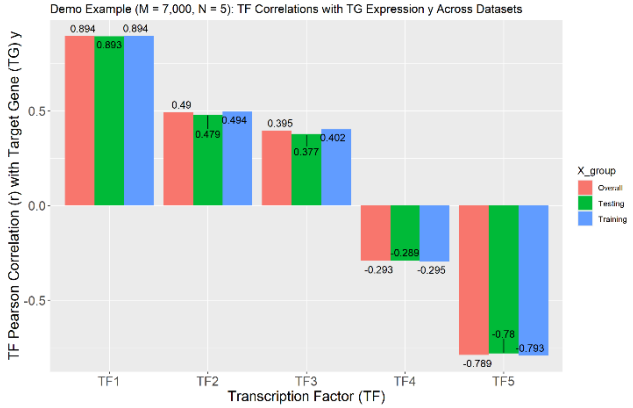


Figure B.1B) - Comparing actual versus expected correlations between predictors and TG y after splitting the data: 70% for training and 30% for testing.

We hold $cor(TF, TG) \approx$ [0.9, 0.5, 0.4, -0.3, -0.8] in the data overall.

Figure B.1C) - Pearson correlation matrix among the TFs and TG in the testing data.

Dot sizes represent magnitude, and colors indicate positive (blue) or negative (red) correlations. These correlations are typically the same before and after standardizing X and centering y. The matrix at the top right (upper triangular) corresponds to the original correlations in test data, while the matrix at bottom left (lower triangular) corresponds to testing data correlations after we standardize expression levels of TFs (X) and TG (y) based on the insights from the training data (e.g. StandardScalar from the training X data and mean of the training y data).

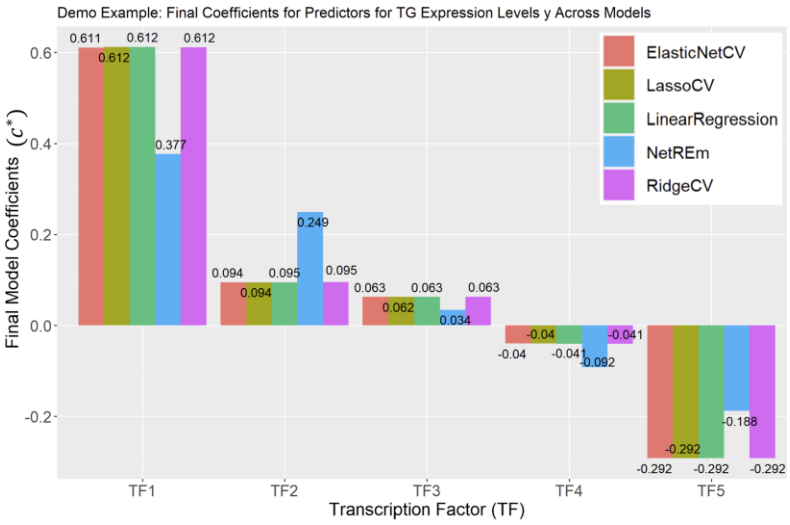


Figure B.1D) - Compare final optimized coefficients c^* predicted by NetREm versus 4 other benchmark regression models (BRMs: LassoCV, Linear Regression, ElasticNetCV, RidgeCV) for same data.

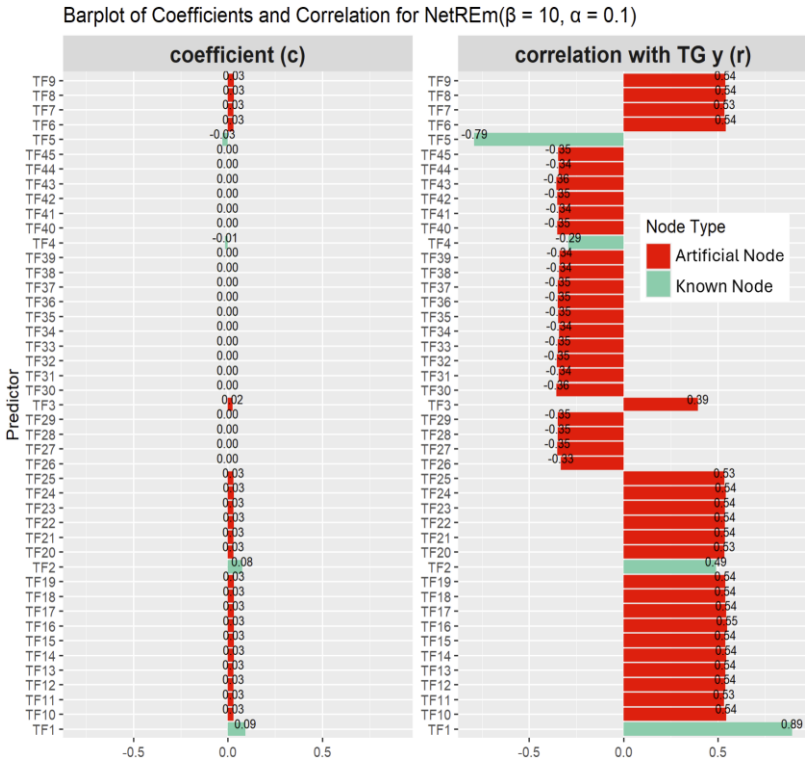


Figure B.1E) – Barplot of coefficients and correlation for NetREm(β = 10, α = 0.1).

We expand **Figure 3.2A** main example for to add 20 more TFs with r = 0.55 with TG, and 20 more TFs with r = -0.35 with TG. Correlations (right plot) are for the training data. We use the **Fig. 3.2A** network of strong known $TF_4 - TF_5$ links and $TF_1 - TF_2$ links (known nodes with PPIs are TF_1, TF_2, TF_4, TF_5) and the rest are artificial nodes and edges (weight = 0.01). So, we have $N = 45$ candidate TFs in this case. We make $\beta = 10$ (and retain $\alpha = 0.1$) and run NetREm for the 7,000 training cells. We illustrate the grouping variable selection property of NetREm, and how TF_1, TF_2, and TF_5 tend to have meaningful coefficients. This is important, since we have added 20 TFs that have a stronger correlation (r = 0.55) with TG y expression levels than TF_2 does and those TFs all end up having lower $|c^*|$ than TF_2 does. Further, TF_4 (has the smallest expression $|r| \approx$ 0.29 with y) is 1 of the $N^* = 25$ TFs that have a non-zero coefficient in c^*. If we used Lasso regression with $\alpha = 0.1$ and no y-intercept (not shown for visual simplicity), we would get $N^* = 23$ TFs, where we lose TF_4 (and TF_3). This shows how NetREm can prioritize TF_4 based on its strong known PPI connection with TF_5.

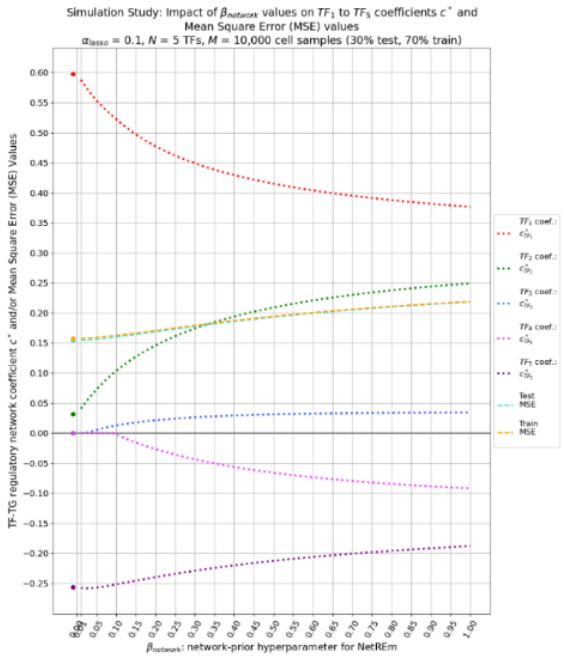


Figure B.1F) – This lineplot explores the effects without network regularization penalty *($\beta = 0$)* and with this network regularization penalty *($0.01 \leq \beta \leq 1$)* on Mean Square Error (MSE) and coefficients.

This visualizes model evaluation MSE results on training and testing data as a function of β. Moreover, it traces impacts of β on coefficients for the 5 TFs, all else constant ($\alpha = 0.1$). As β increases: train and test MSE values increase, the magnitude of coefficients for TF_1 and TF_5 decrease, and magnitudes of coefficients for TF_2 and TF_4 increase much greater than for TF_3. NetREm(β = 0.01, LassoCV α) achieves test MSE ≈0.14.

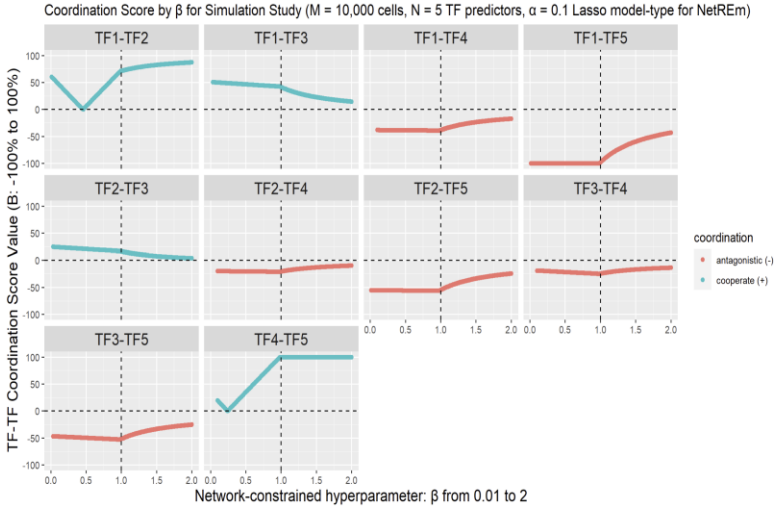


Figure B.1G) – Coordination Score by β for Simulation Study (M = 10,000 cells, N = 5 TF predictors, α = 0.1 Lasso model-type for NetREm)

For the main simulation study in **Fig. 3.2A**, we vary β from 0.01 to 2 and analyze the changes in TF-TF coordination scores, when we use the Lasso NetREm model type with $\alpha = 0.1$. Cooperative (+) relationships are in red and antagonistic (-) relationships are in blue. The larger β is, the more confidence we have in the input PPIN; we want the input PPIN to then carry more weight in terms of deciding outcomes by grouping predictors.

Figure B.2 Adapting Simulation Study for Various Sparsity (%) of the Underlying Data

Rerunning the analysis in **Fig. 3.2A-C** (40% sparsity) for 4 different sparsity levels: 10%, 50%, 60%, 70%. NetREm models are run with same settings: no y-intercept, $\alpha = 0.1$, $\beta = 1$ and tend to yield the same results as those in **Fig. 3.2** in the manuscript.

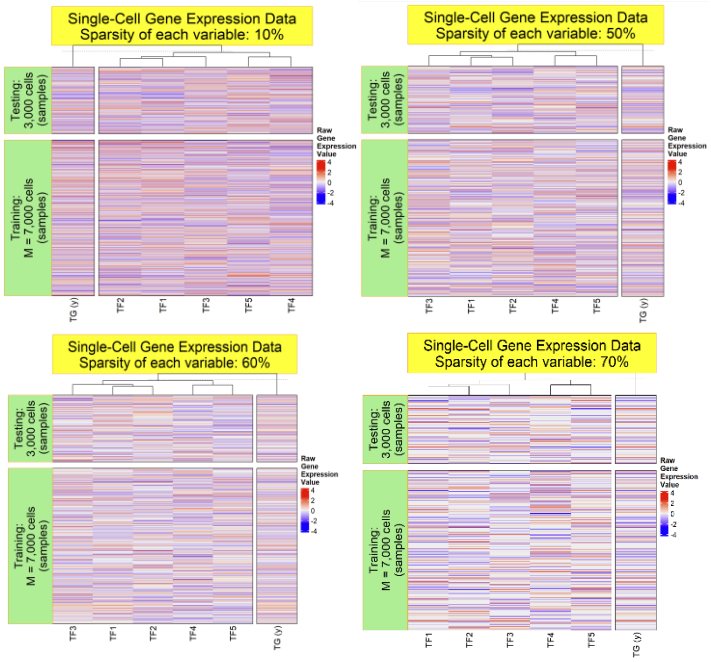


Figure B.2A) – Gene expression data for the respective simulations and sparsity levels.

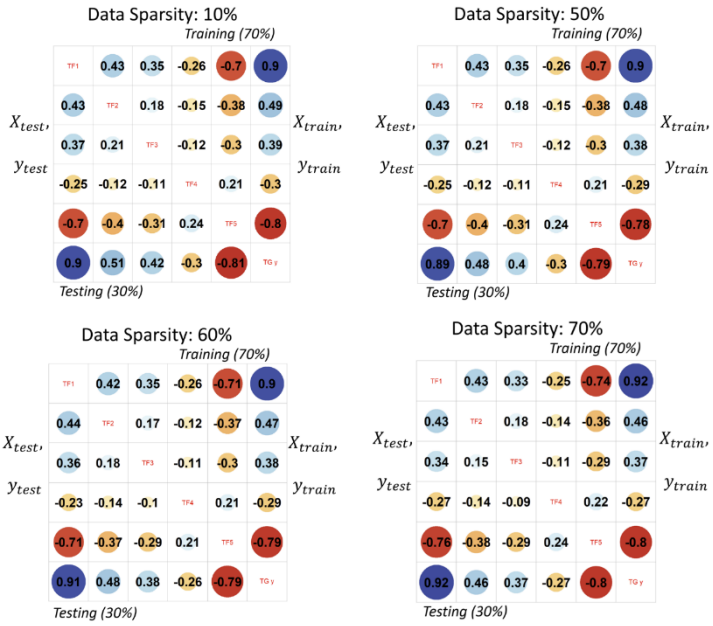


Figure B.2B) – Corresponding correlations in training and testing data sets for the respective sparsity %.

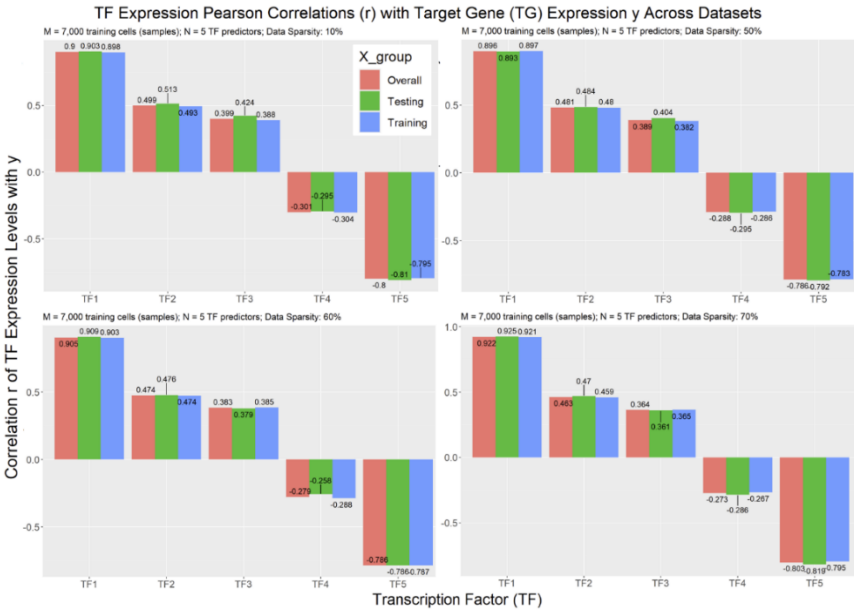


Figure B.2C) – Expected versus Actual correlations of each TF predictor's expression level with Target Gene (TG) expression levels, after incorporating sparsity and partitioning data into train and test data.

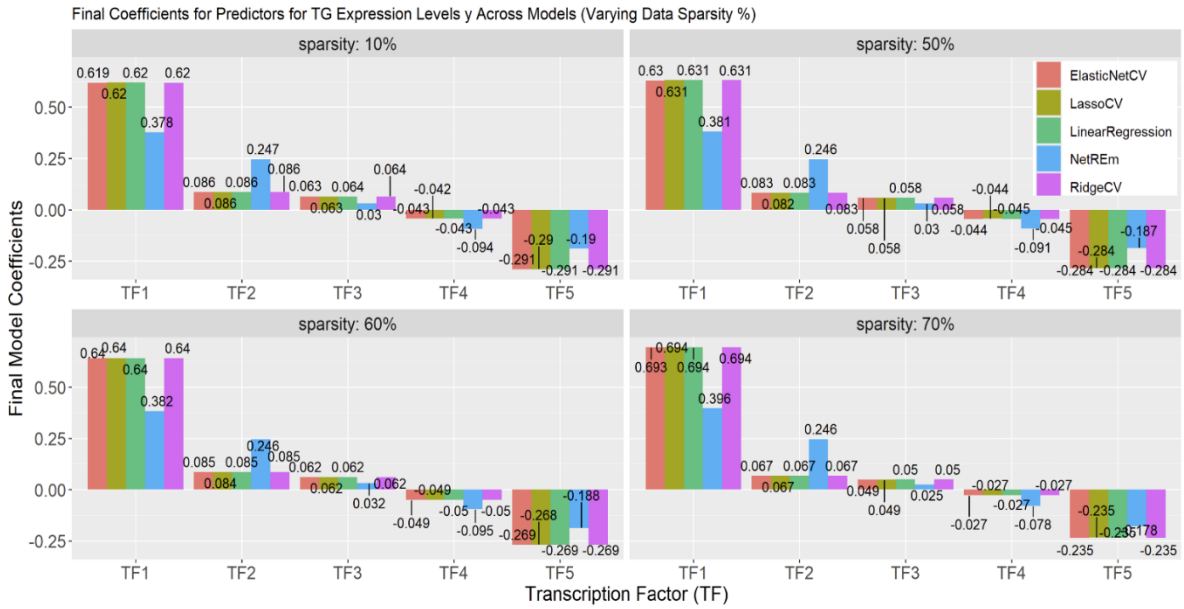


Figure B.2D) – Final coefficients c^* for TFs for predicting TG expression y. Results based on sparsity % of the underlying data.

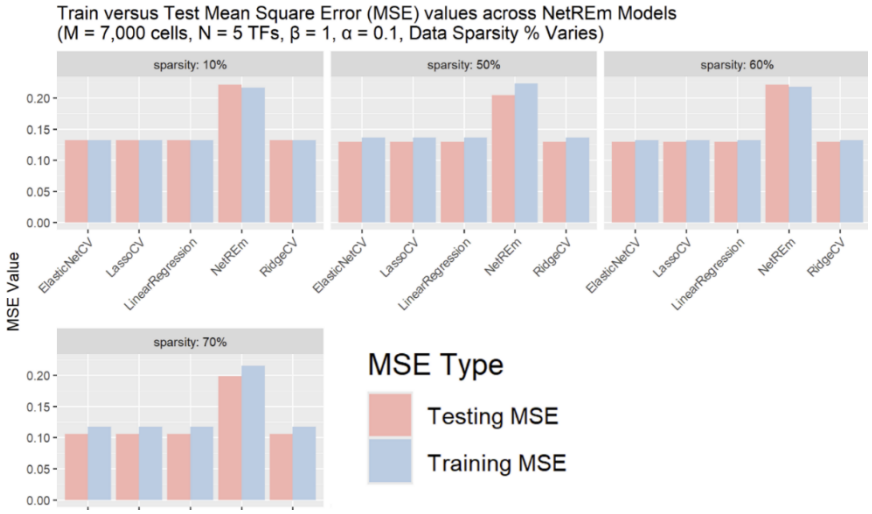


Figure B.2E) – Mean Square Error (MSE) on train (70%: 7,000 cells) versus test data (30%: 3,000 cells).

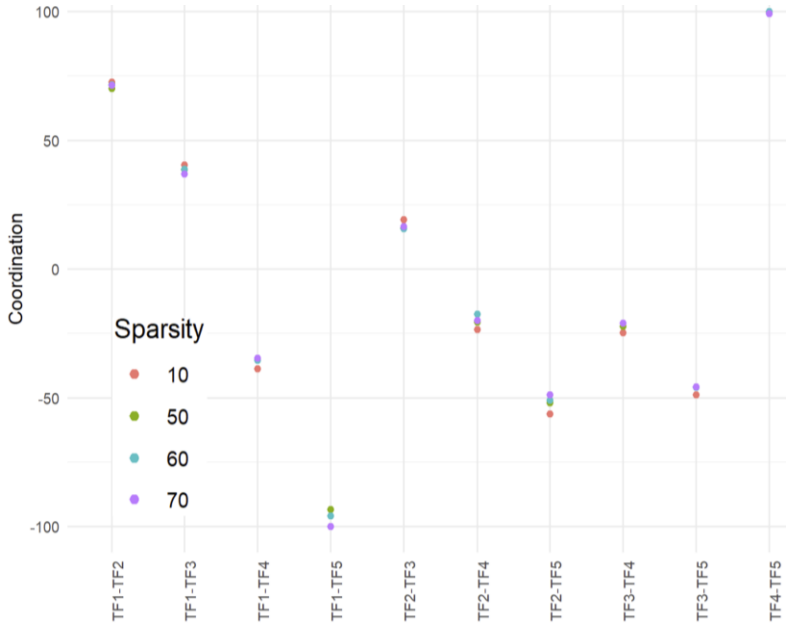


Figure B.2F) – TF-TF coordination scores for regulating the Target Gene (TG).

Figure B.3 Analysis of NetREm performance based on 1,000 simulations

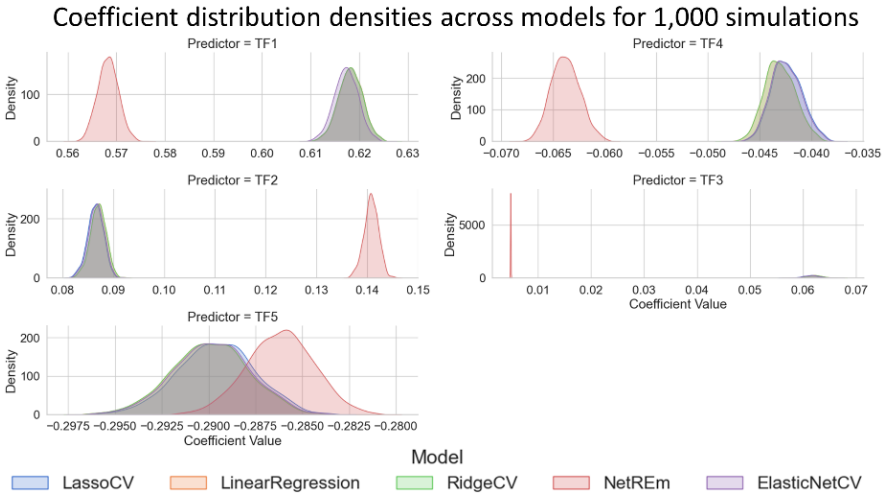


Figure B.3A) - Density distribution of coefficients for the 5 TFs across NetREm and the 4 benchmark models (LassoCV, LinearRegression, RidgeCV, ElasticNetCV) for 1,000 simulations with the same $cor(TF, TG)$ and prior biological network.

In particular, coefficients for TF_1, TF_2, TF_5 typically display exceptionally low standard deviation (std), underscoring NetREm's stability and consistent inference of minimal influence in the model. The coefficient density distribution plot highlights narrower and more peaked distributions of NetREm's coefficients, indicating reduced variability lower spread, enhanced stability across simulations. Overall, NetREm enhances robustness via network constraint regularization, proving particularly advantageous in high-dimensional data prone to multicollinearity and overfitting. NetREm's density distributions of c^* are more peaked.

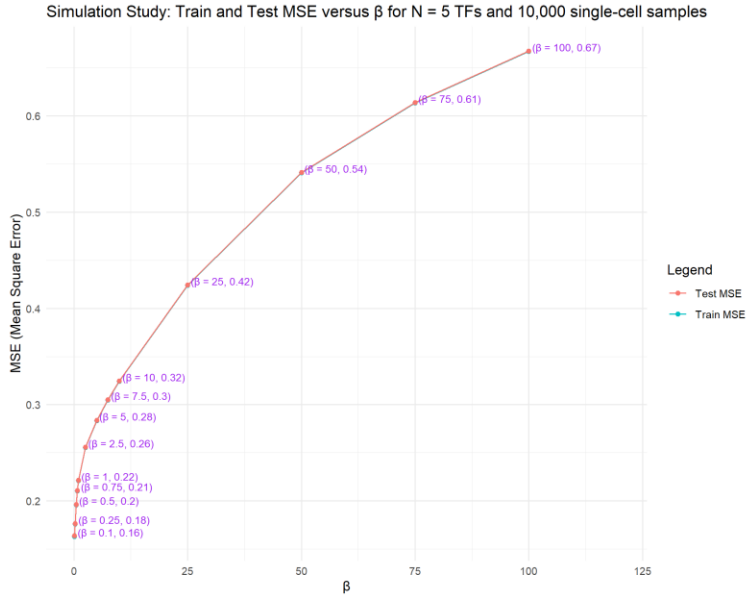


Figure B.3B) - Line plot of average training and testing Mean Square Error (MSE) values across the 1,000 simulations, for 13 different NetREm models on the same input data (benchmarking β).

Each model is based on a varied network constrained hyperparameter β. Here, as β increases, the MSE values tend to increase monotonically.

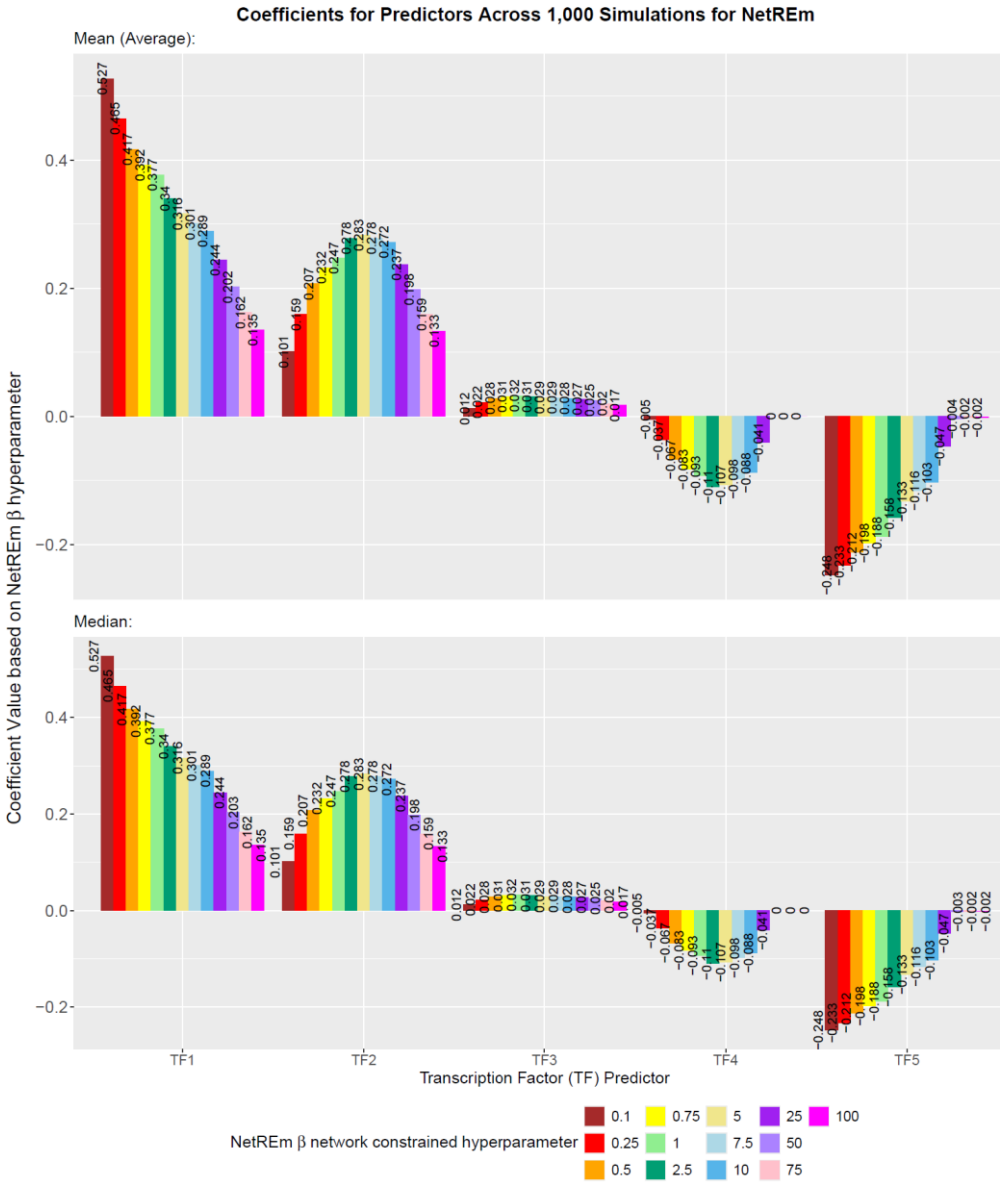


Figure B.3C) - Comparative barplots of the average and median coefficients for the 5 TF predictors across the 1,000 simulations for the 13 NetREm models (varying network constrained hyperparameter β). That is, this is additional analysis based on the benchmark β results in **Figure B.3B**. We note that for $\beta > 5$, the network constraint is too harsh. This excessive network regularization penalty can be detrimental, leading to models overly constrained by the predefined network structure as the magnitude of coefficients tends to decrease. In particular, the magnitude of coefficients for TF_2 and TF_4 is a U-shape, which increases for $\beta \in \{0.1, 0.25, 0.5, 0.75, 1, 2.5, 5, 7.5, 10, 25, 50, 75, 100\}$ and starts decreasing for $\beta \in \{7.5, 10, 25, 50, 75, 100\}$. By seamlessly blending predictive precision with structural intelligibility, NetREm stands out for its superior consistency, potential to unveil complex data structures, and computational efficacy.

Figure B.4 Simulations and theoretical analysis of various cases relating M to N and associated singular values (s). (Under-the-hood analysis).

Singular values (s) of E matrix for simulations comparing M: the # of single-cell samples (cells) with N: the # of TF predictors for X matrix where $X \in \mathbb{R}^{M \times N}$. This figure showcases examples where $N \ll M$, $N = M$, or $N \gg M$, to illustrate how NetREm can tackle different scenarios without having the resulting E matrix become ill-conditioned. That is, $\kappa(E) \ll$ 1e6 in our examples (meeting criteria of being relatively well-conditioned). We incorporate sparsity so each gene (predictors and TG) has ~same % sparsity: 40% for **Fig. B.4A-E**, 70% for **Fig. B.4F**. We use $\beta = 1$ and $\alpha = 0.1$. The following are illustrative examples of results we retrieve from running NetREm on our simulated and standardized X and y data. We randomize the edge list weights for the prior network in each simulation for **Fig. B.4B** to **Fig. B.4H. Fig. B.4F)** and **Fig. B.4E)** have the same N and M values and show the impact of sparsity since **Fig. B.4E)** has 40% sparse data and **Fig. B.4F)** has 70% sparse data. In these examples **Fig. B.4A-F**, we note the range of largest $s_{\max}$ to smallest $s_{\min}$ singular value is well-contained. All of the $\kappa(E)$ values are much smaller than 1e6. **Figures B.4G** and **B.4H** focus on the situation where we fix N = 50 TFs and then adjust the number of samples M from 2 to 200. We show the transition from $M \ll N$ to $M < N$ to $M = N$ (denoted by a blue dashed line) to $M > N$. Thus, we are transitioning from where N is ~25M to where M is 4N. $\kappa(E) \approx 8{,}382.36$ and converges at a horizontal asymptote ~ 20 to 40.

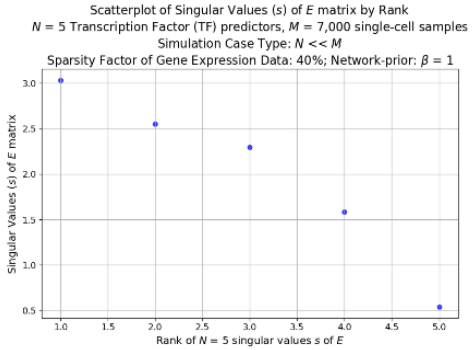


Figure B.4A) $N = 5$ singular values ($\mathcal{S}$) of E matrix (upon which SVD is performed) for data in **Fig. 3.2A** demo shown in the simulation study.

Here $M = 7{,}000$ training samples so ($N \ll M$): N = 5, M = 7,000. Then, $\mathcal{S}$ = [3.033, 2.548, 2.296, 1.583, 0.54] and the resulting condition # for E is: $\kappa(E) = \frac{s_{\max}}{s_{\min}} = \frac{3.03296279}{0.53976039} \approx 5.619091$.

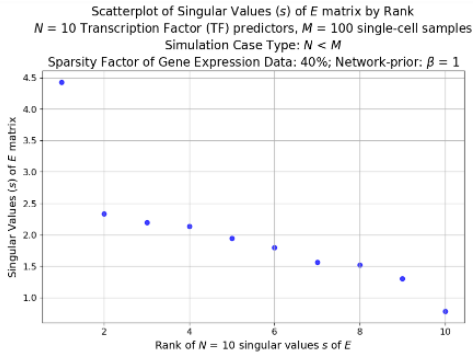


Figure B.4B ($N < M$): N = 10, M = 100→ $s_{max} = 0.783$ and $s_{min} = 0.1767$. Then, $\kappa(E)$ is 5.

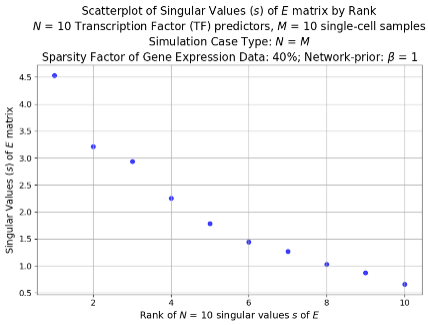


Figure B.4C ($N = M$)*: N = 10, M = 10* → $s_{max} = 0.663$ and $s_{min} = 0.146,$ so $\kappa(E)$ is 6.84.

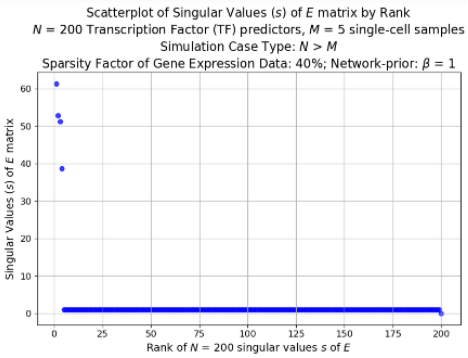


Figure B.4D ($N > M$): *N = 200, M = 5* → $s_{max} = 2.07e-2,\ s_{min} = 3.387e-4$ → $\kappa(E)$ is 2,952.29.

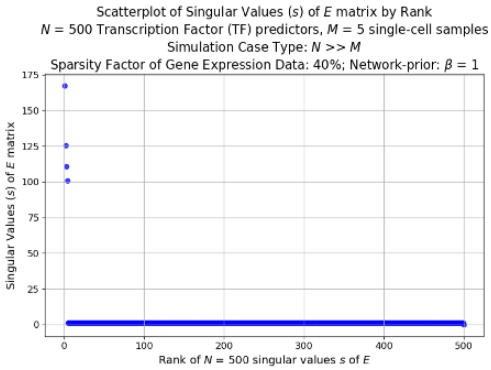


Figure B.4E ($N \gg M$)*: N* = 500, *M* = 5→ $s_{max} = 9.04e-3, s_{min} = 5.3989e-5 \rightarrow \kappa(E)$ is 18,522.4.

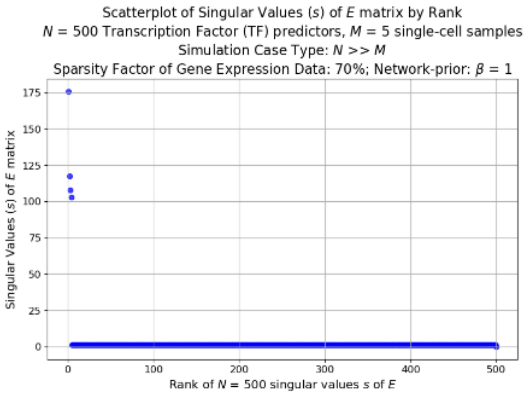


Figure B.4F ($N \gg M$): *N* = 500, *M* = 5 → $s_{max} = 1.57e-2,\ s_{min} = 8.96e-5$ → $\kappa(E)$ is 11,152.384.

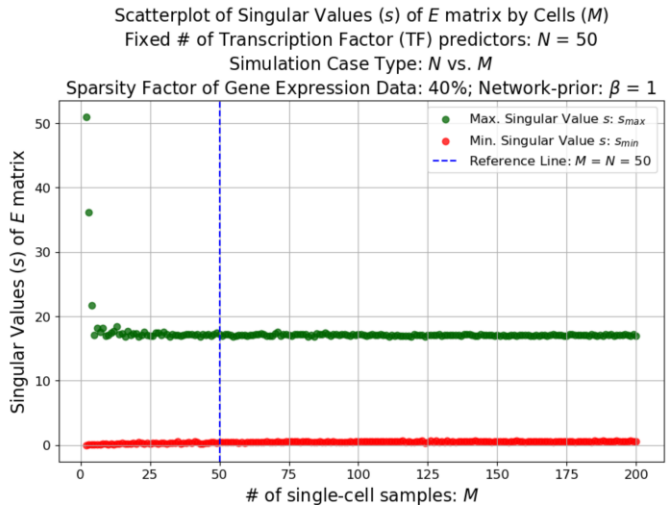


Figure. B.4G shows how the maximum s_{max} and minimum s_{min} singular values evolve during this continuous transition.

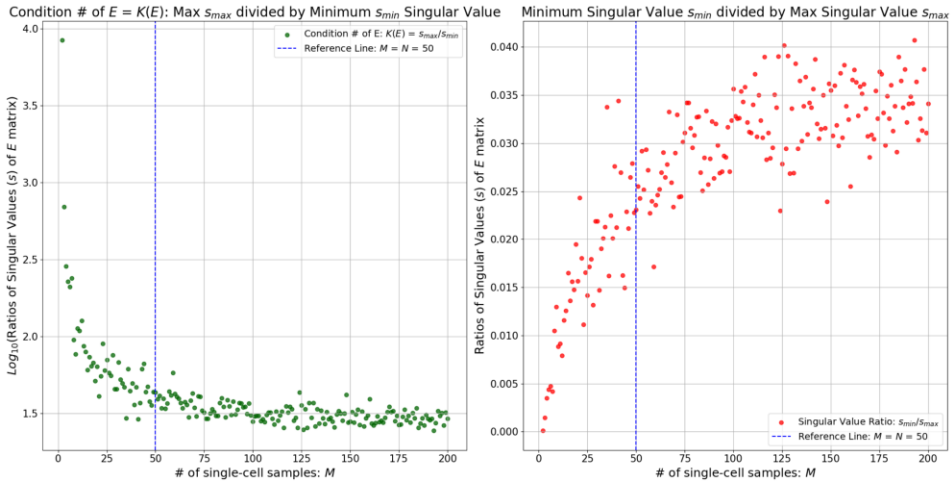


Figure B.4H has 2 plots. 1 shows ratio of condition number of E, $\kappa(E) = \frac{s_{max}}{s_{min}}$ and the other shows reverse ratio (s_{min}/s_{max}). For instance, when M is 2 and N is 50, $\kappa(E)$ is 8,382.36 << 1e6. Thus, we anticipate NetREm can be applied to any foreseeable scenario.

Figure B.5 Simulation Study adapted for a case where predictors $N > M$ cells (samples).

We have 6 TF predictors $[TF_1, TF_2, TF_3, TF_4, TF_5, TF_6]$ for the TG and $M = 5$ cells (single-cell samples) and use all data for training. These 6 TFs have respective expression levels $X \in \mathbb{R}^{5 \times 6}$ and TG has expression levels $y = [y_1 \quad y_2 \quad y_3 \quad y_4 \quad y_5]^T$. **Figure B.5F-I** focus on outputs of NetREm run with network regularization hyperparameter $\beta = 1$ and sparsity hyperparameter $\alpha = 0.1$.

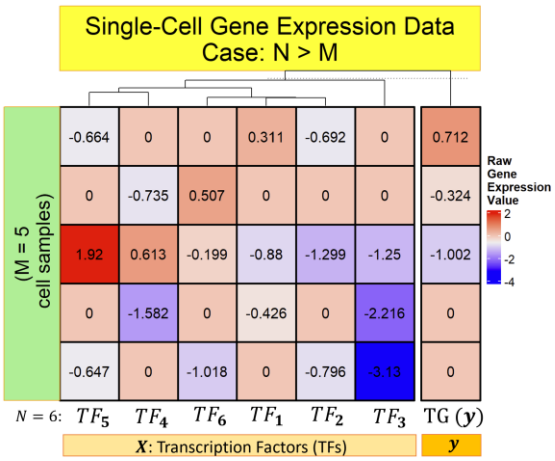


Figure B.5A) Single-cell gene expression data.

We ensure 40% sparsity for each variable, so 2 cells out of 5 for each TF are 0. These are expression values prior to standardization. Ultimately, we standardize this expression data so that each variable has a mean of 0 and standard deviation of 1.

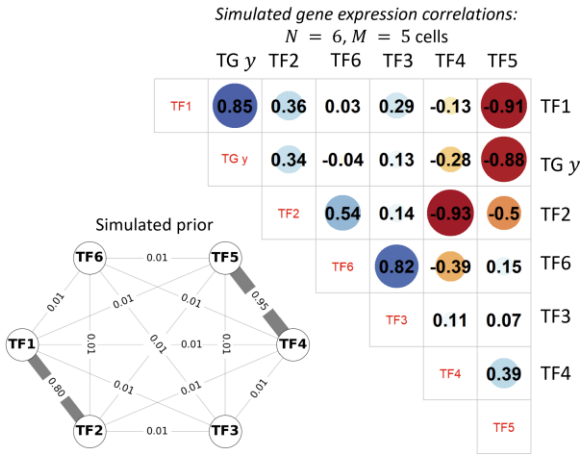


Figure B.5B) Bottom left shows the prior biological network with default edge weights (0.1), with stronger experimentally-verified connections for TF_1- TF_2 *(0.8)* and TF_4- TF_5 *(0.95)*.

The top right presents a correlation r matrix among TFs and *TG* in the training expression data where: $cor(TF, TG) \approx [0.85, 0.36, 0.03, 0.29, -0.13, -0.91]$ for TF_1 to TF_6. Dot sizes represent magnitude, and colors indicate $r > 0$ (blue) or $r < 0$ (red)

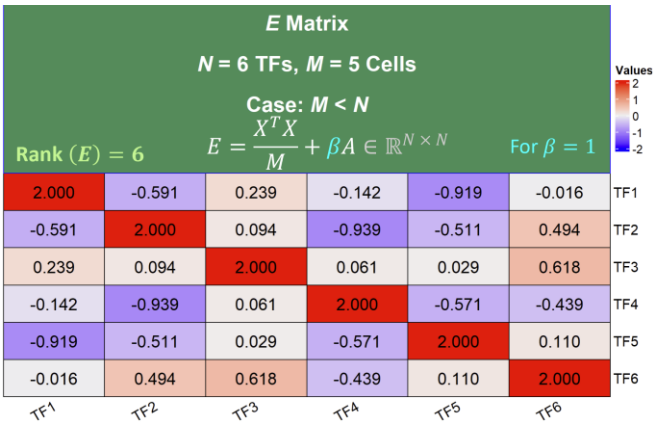


Figure B.5C) Breakdown of values in symmetric, positive semi-definite matrix $E = \frac{X^T X}{M} + \beta A \in \mathbb{R}^{N \times N}$.

Since we select $\beta = 1$, we have $E = \frac{X^T X}{M} + A$. We display values of E, noting E is of full rank since it is 6 and rank(E) can be a max of $\min(\text{rows} = 6,\ \text{columns} = 6) = 6$.

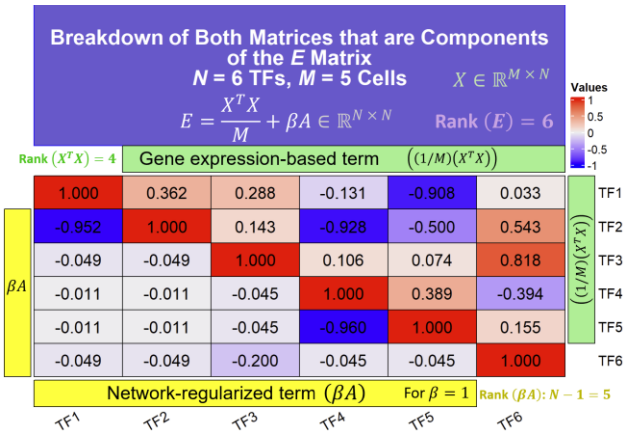


Figure B.5D) Individual components of E.

The 1st term (Gene expression-based term $\frac{X^T X}{M}$), is on top right (above main diagonal of all 1). This matrix always has main diagonal of 1 since expression data is standardized. This matrix has a rank of 4 in this case, and this Gram matrix $X^T X$ thus is not full rank since max rank that $X \in \mathbb{R}^{5 \times 6}$ can be $\min(M = 5,\ N = 6) = 5$. The 2nd term (network-regularized term βA) is on bottom left (below main diagonal). When $\beta = 1$, this 2nd term has a main diagonal of 1 as well. Since A is a variant of the Graph Laplacian matrix, it has rank $N - 1$, which is $6 - 1 = 5$. Hence, main diagonal of E is 2. Ultimately, both terms help E achieve a full rank, which we show in **Figure B.5D**.

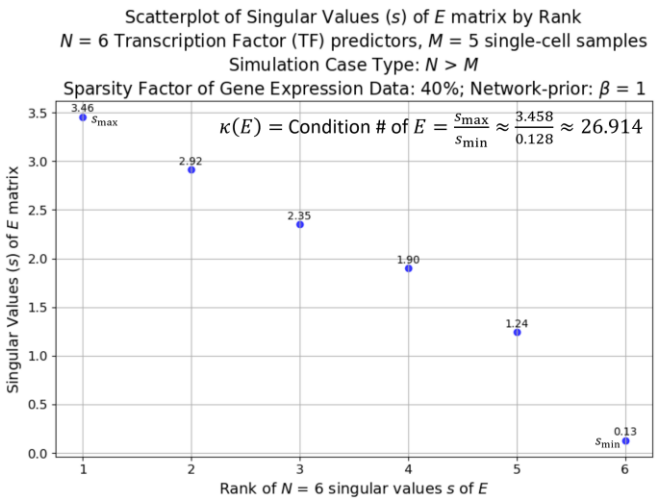


Figure B.5E) NetREm performs Singular Value Decomposition (SVD) on E to uncover 6 positive singular values: $\mathcal{S} = \{s_1, s_2, s_3, s_4, s_5, s_6\}$). We find that $s_{min} = 0.1284$, $s_{max} = 3.457$, $\frac{s_{min}}{s_{max}} \approx 0.0372$, and the condition # of $E = \kappa(E) = \frac{s_{max}}{s_{min}} = 26.9136 \ll 1e6$, so E is well-conditioned.

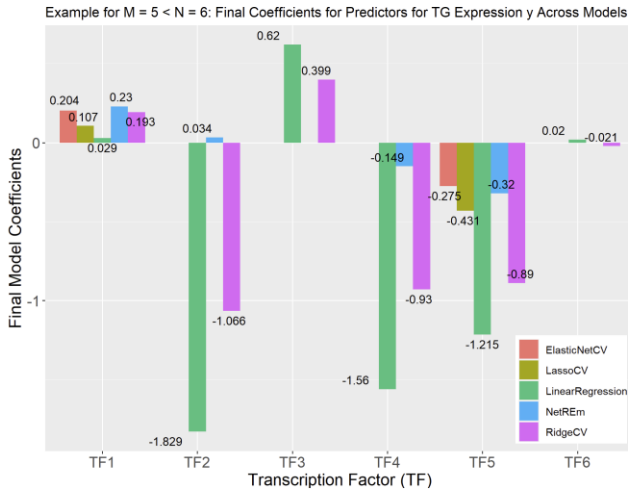


Figure B.5F) Final coefficients c^* for each TF predictor for regulating TG in a TF-TG regulatory network. NetREm coefficients compared to those from the other 4 benchmark regression models.

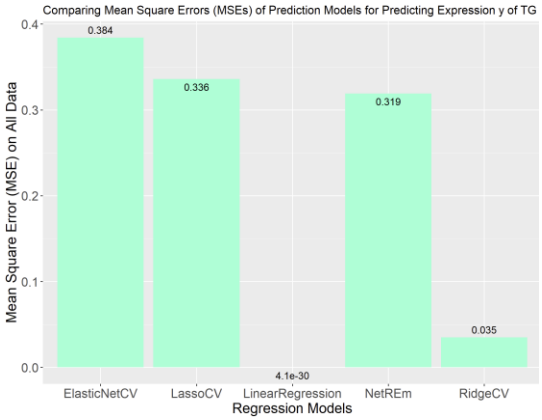


Figure B.5G) Mean square error (MSE) values for 5 models.

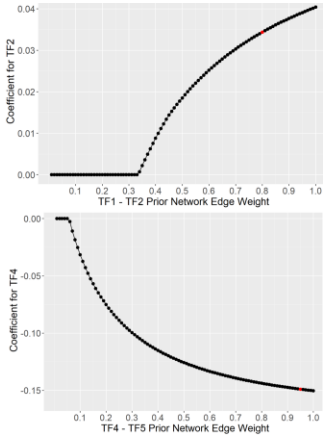


Figure B.5H) Top: Effects of varying TF_1-TF_2 edge weight in original simulated biological prior network, from 0.01 to 1 in 0.01 increments, holding all other parameters fixed (e.g. Lasso model for NetREm with $\alpha = 0.1$*)*. Respective TF_2 c^* increases monotonically in arc shape. Red dot: selected edge weight used for simulated prior network. Bottom: Similar sensitivity analysis for TF_4 and TF_5 shows TF_4 $|c^*|$ increases, becoming more negative from 0, as the TF_4-TF_5 edge weight is perturbed.

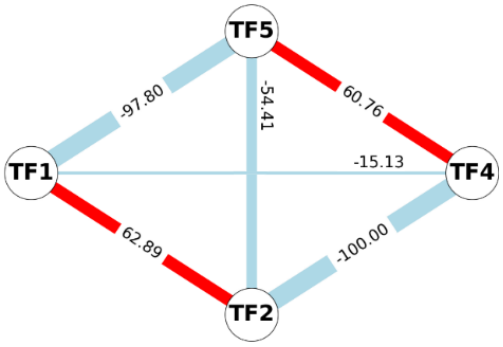


Figure B.5I) TF-TF coordination network for the TG. Red: cooperative ($B > 0$) interactions. Blue: antagonistic *($B < 0$)* interactions between the TFs. $-100 \leq B \leq 100$.

Figure B.6 Evaluating TF-TG Regulatory Network properties in simulated Human Embryonic Stem Cells (hESCs)

We benchmark NetREm's performance for constructing TF-TG regulatory links for 1,250 TGs in hESCs across each of the 6 SERGIO-simulated datasets, which vary based on noise parameter (30%, 60%, 90%) and training cells M (70 versus 700) based on 70% of data being used for training. SERGIO incorporates TF-TG regulatory interactions (e.g. stochastic nature of transcription, regulation of TFs by many TFs) found in true data. Data simulated by SERGIO is comparable to experimental data generated by Illumina HiSeq2000, Drop-seq, Illumina 10X chromium, Smart-seq. We input the comprehensive human PPIN of 21,321 edges comprising 10,777 known links and 10,544 artificial links (η = 0.01) for these 207 TFs. For each of the 15 TFs that are also TGs, we remove respective TF and its associated edges from input PPIN and note $N = 206$ for them; $N = \mathcal{N}$= 207 for remaining 1,235 TGs. We run 6 NetREm models with $\beta = 1$, varying α (0.01, 0.025, 0.05, 0.075, 0.1, LassoCV). As α increases, # of discovered TF-TG links decreases as expected, given stricter α leads to sparser results (i.e. fewer $c^* \neq 0$ across TGs).

Figure B.6A) Focuses on 5 metrics (Balanced Accuracy (BACC), F1 scores, Precision, Sensitivity, Specificity) of SERGIO-simulated data where $N < M = 700$ cells in train data.

As α increases, specificity, precision, and F1 scores improve; balanced accuracy (BACC) and sensitivity decline, particularly for $M = 700$. For 90% noise, NetREm models maintain or exceed a precision of 2 to 2.5% compared to benchmarks ≈2%.

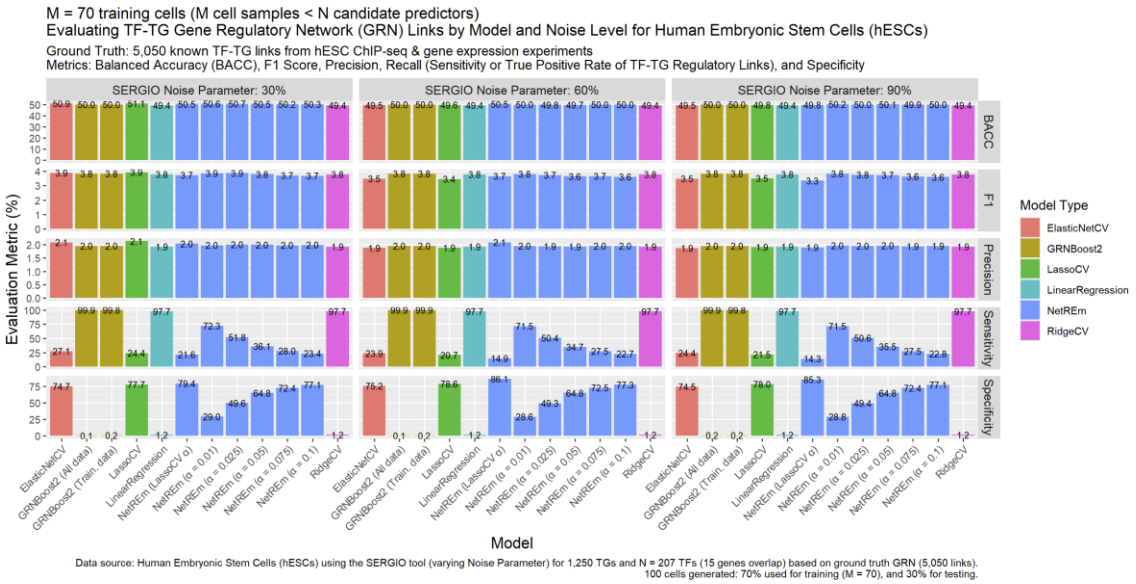


Figure B.6B) 5 metrics for data where $N > M = 70$ cells in train data. When $M = 70$, increased α results in higher specificity but lower sensitivity and general decreases in F1 scores.

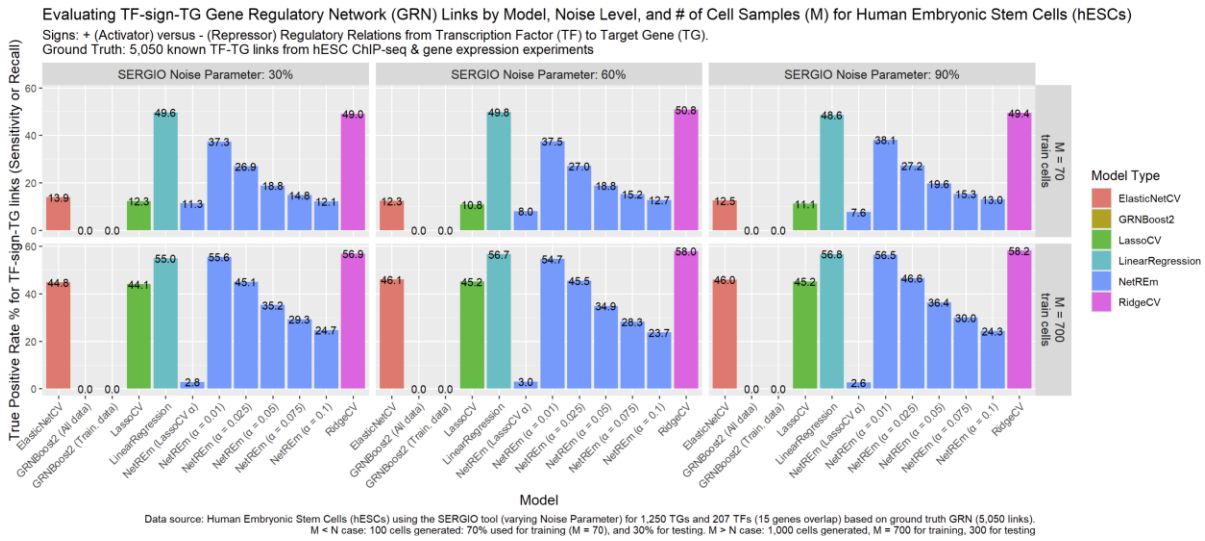


Figure B.6C) True positive (TP) of TF-sign-TG links where sign ($c^* > 0$*: activator,* $c^* < 0$*: repressor*) matters.

Comparing results across noise for 2 values of M. We compare NetREm with BRMs in terms of signed TF-TG links, focusing on predicting c^* signs (+: activate, -: repress) (**Fig. B.6C-D**).

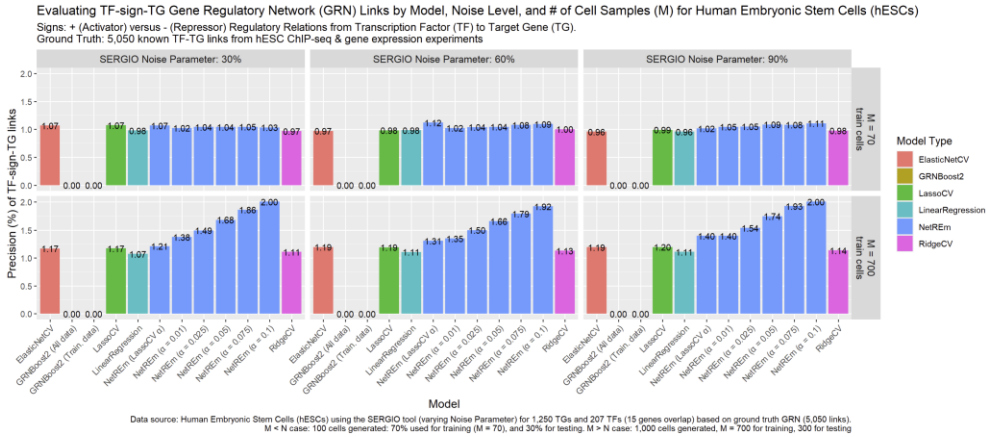


Figure B.6D) Focuses on Precision of TF-sign-TG regulatory links.

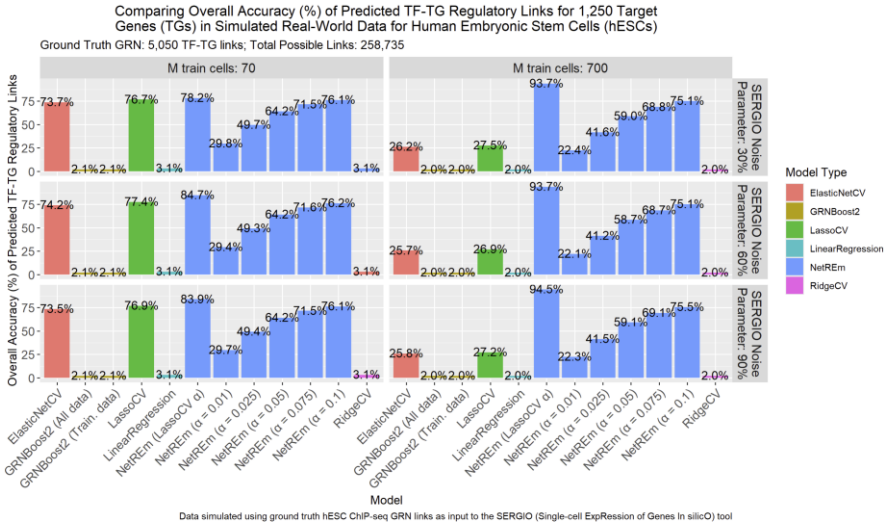


Figure B.6E) Comparing accuracy (TPs + True Negatives (TNs))/(Total predictions).

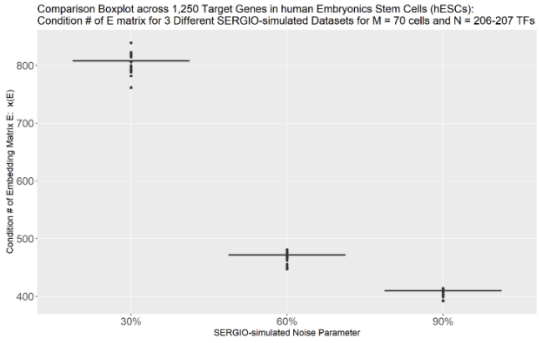


Figure B.6F) Boxplots of max condition # of E across all 1,250 TGs for $M = 70$. For 30% noise: $max(\kappa(E)) = 838.9$, for 60% noise: $max(\kappa(E)) = 480.91$, for 90% noise: $max(\kappa(E)) = 414.12$. In an additional sanity check, for $M = 70$, we find max $\kappa(E) < 900$ across TGs for different noise levels.

Figure B.7 Highlighting NetREm's Grouped TF Selection Property (prioritizing TF-TF relations in PPIN) using 4 Target Genes (TGs) in human Hematopoietic Stem Cells (HSCs)

This figure highlights NetREm's potential to flag groups of coordinating TFs connected along biologically meaningful, cell-type-specific PPIs (Li and Li 2008). Novel TFs: not in the ground truth GRN for that TG in HSCs. **Figures B.7A-F** focus on NetREm's results for *ATF2*: NetREm($\beta = 10$, LassoCV to identify optimal α) identifies $N^* = 8$ final TFs for *ATF2* regulation in HSCs, and 7 of the 8 (excluding WHSC1) are substantiated in ground truth validation GRNs (Zhang et al. 2023) for *ATF2* in HSCs. **Figures B.7G-I**: enriched PPIs among final TFs for other target genes (TGs): *DUSP2, BRD2, RNF167*. These are examples of NetREm uncovering meaningful subnetworks of PPIs among predictors for 3 other TGs (*BRD2*: 13 of 21 TFs are validated, RNF167: 17 of 27 TFs are validated, *DUPS2*: 25 of 47 TFs are validated) missing in ElasticNet and Lasso models and predicted to have all candidate TFs by Ridge and Linear regression models.

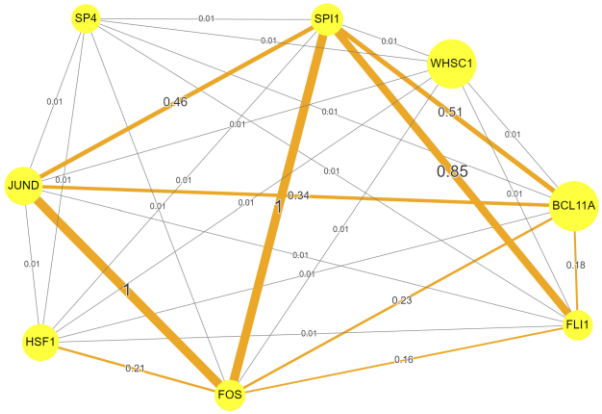


Figure B.7A) Comprehensive input PPI network (PPIN) among these 8 final TFs for ATF2.

There are strong known PPIs between FOS-SPI1 and FOS-JUND ($w = 1$) and beyond. Here, TFs are yellow nodes and orange links (with w) are input PPIN weights between TFs. Grey links are artificial (default weight: $\eta = 0.01$). SP4 and WHSC1 lack documented PPIs in this PPIN. NetREm performs grouped variable selection of TFs based on a fully-connected input PPIN of known PPIs and artificially-added weak links.

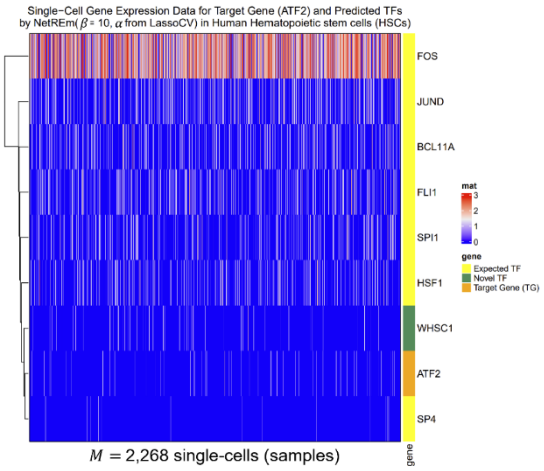


Figure B.7B) Gene expression levels for SP4 and WHSC1 hierarchically cluster closest to those of ATF2 in HSCs (based on dendrogram on the left), which may explain their inclusion as final TFs.

Pairwise correlations (r) Among TFs and the Target Gene (TG) ATF2 in Human Hematopoietic Stem Cells									
	BCL11A	ATF2	WHSC1	SP4	HSF1	SPI1	FLI1	FOS	JUND
BCL11A		0.004	-0.047	-0.006	0.009	0.009	-0.027	-0.027	-0.047
ATF2	0.004		0.002	0.009	-0.032	0.020	0.025	0.004	0.016
WHSC1	-0.047	0.002		-0.017	0.016	-0.008	0.013	0.003	-0.004
SP4	-0.006	0.009	-0.017		0.050	0.047	-0.020	0.016	-0.019
HSF1	0.009	-0.032	0.016	0.050		0.023	-0.005	0.011	-0.042
SPI1	0.009	0.020	-0.008	0.047	0.023		0.022	-0.003	-0.008
FLI1	-0.027	0.025	0.013	-0.020	-0.005	0.022		0.031	-0.008
FOS	-0.027	0.004	0.003	0.016	0.011	-0.003	0.031		0.045
JUND	-0.047	0.016	-0.004	-0.019	-0.042	-0.008	-0.008	0.045	

Figure B.7C) Pearson Correlations *(r)* among 8 TFs and TG ATF2 in HSCs across 2,268 cells.

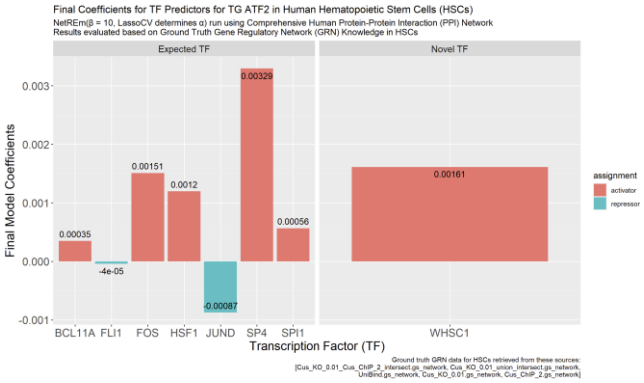


Figure B.7D) TF-TG regulatory network returned by NetREm for *ATF2* where coefficients $c^* > 0$ for activator TFs and $c^* < 0$ for repressor TFs.

Literature supports FLI1 (core TF for hematopoiesis) and JUND may act as repressors (Chen et al. 2008; Liu et al. 2019) and others as activators.

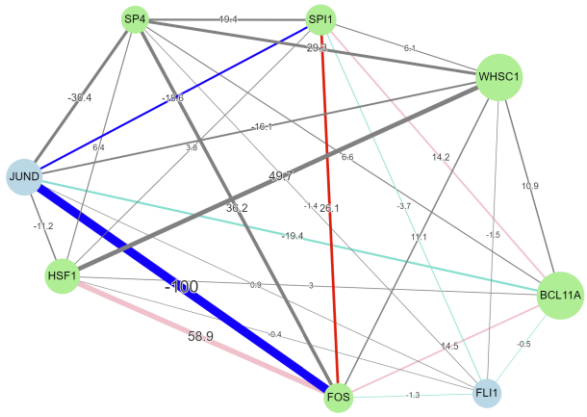


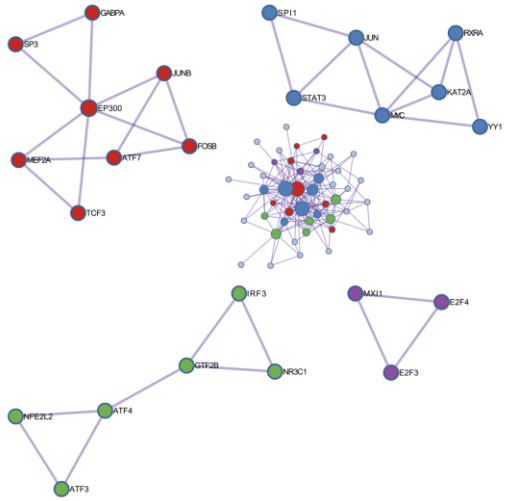
Figure B.7E) NetREm outputs ATF2-specific TF-TF coordination network for the 8 TFs.

Red = + relationship (cooperative) and found in input PPIN and PPI is associated with immune-related conditions in contextual PPI database (PPID) (Kotlyar et. al 2022). Pink = + relationship (cooperative) and known PPI. Blue = - relationship (antagonistic) and known PPI (in input PPIn) and associated with immune-related conditions in contextual PPID. Turquoise = - relationship (antagonistic) and found in input PPIN. Grey = Novel relationship (not found in the comprehensive input PPIN). Contextual PPID corroborates antagonistic JUND-FOS and JUND-SPI and cooperative SPI1-FOS links are immune-related (e.g. bone marrow, lymph nodes, synovial macrophages, leukemia, multiple myeloma, lymphoma, immune system cancer, bone inflammation disease, myeloid leukemia, bone marrow cancer, myeloid neoplasm). NetREm uncovers potential antagonism between SPI1 and FLI1; SPI1 has an inhibitory effect on FLI1-induced erythroid progenitor dedifferentiation (Vecchiarelli-Federico et al. 2017).

GO	Description	Log10(P)
WP2849	Hematopoietic stem cell differentiation	-7.6
hsa04380	Osteoclast differentiation	-6.5
GO:0030099	myeloid cell differentiation	-5.5

Figure B.7F) Metascape *(Zhou et al. 2019b)* shows sequential JUND-FOS-SPI1- FLI1 PPIs are strongly enriched for roles in HSC differentiation among others.

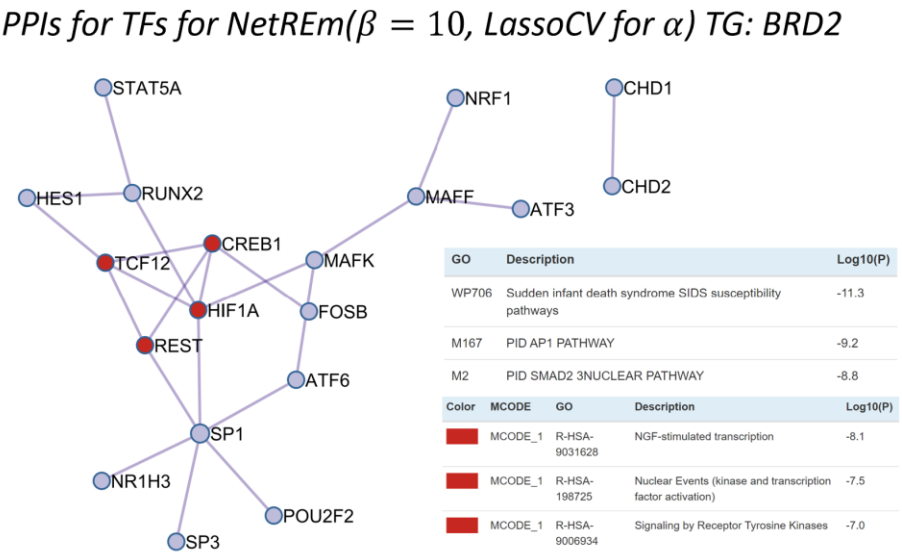
PPIs for TFs for NetREm($\beta = 10$, LassoCV for α) TG: DUSP2



Color	MCODE	GO	Description	Log10(P)
	MCODE_1	R-HSA-198725	Nuclear Events (kinase and transcription factor activation)	-9.0
	MCODE_1	R-HSA-187037	Signaling by NTRK1 (TRKA)	-7.9
	MCODE_1	WP366	TGF beta signaling pathway	-7.6
	MCODE_2	GO:1902893	regulation of miRNA transcription	-11.8
	MCODE_2	GO:2000628	regulation of miRNA metabolic process	-11.3
	MCODE_2	GO:0140747	regulation of ncRNA transcription	-11.2
	MCODE_3	WP3613	Photodynamic therapy induced unfolded protein response	-7.9
	MCODE_3	GO:0030968	endoplasmic reticulum unfolded protein response	-7.0
	MCODE_3	GO:0034620	cellular response to unfolded protein	-6.6

Figure B.7G) Metascape-enriched PPIs among the 47 predicted TFs for *DUSP2*.

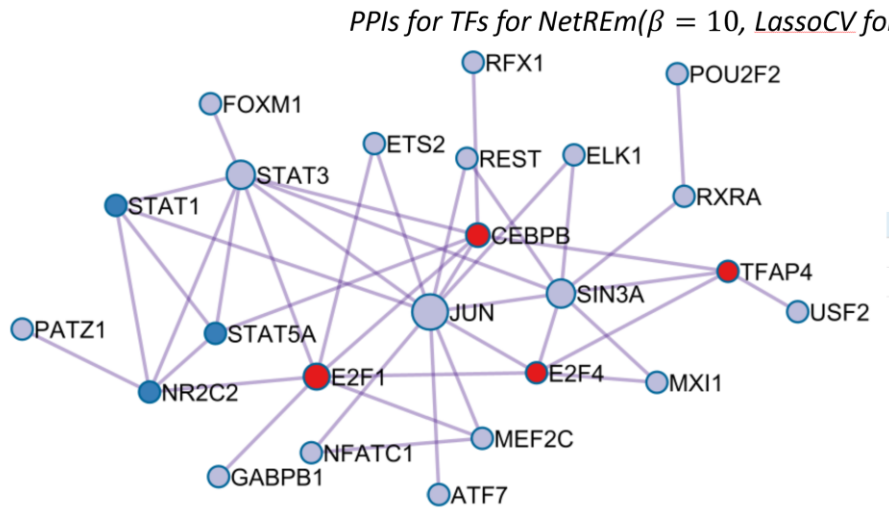
Novel TFs: ATF4, ATF5, CLOCK, E2F3, ERG, FOSB, GTF2B, IRF3, JUN, KAT2A, KLF13, NFE2, NFE2L2, NFE2L3, NFYB, NR2F6, NR3C1, REL, RELB, SKIL, STAT2, STAT3. In *DUSP2*, FOSB is involved in the TGF beta signaling pathway with 7 substantiated TFs, ATF4 and NFE2L2 involved in cellular response to unfolded protein with 5 substantiated TFs, and JUN, KAT2A, STAT3 are involved in regulating miRNA and ncRNA transcription and metabolic processes with other substantiated TFs.



GO	Description	Log10(P)
WP706	Sudden infant death syndrome SIDS susceptibility pathways	-11.3
M167	PID AP1 PATHWAY	-9.2
M2	PID SMAD2 3NUCLEAR PATHWAY	-8.8

Color	MCODE	GO	Description	Log10(P)
	MCODE_1	R-HSA-9031628	NGF-stimulated transcription	-8.1
	MCODE_1	R-HSA-198725	Nuclear Events (kinase and transcription factor activation)	-7.5
	MCODE_1	R-HSA-9006934	Signaling by Receptor Tyrosine Kinases	-7.0

Figure B.7H) Metascape-enriched PPIs among 21 predicted TFs for *BRD2*.
Novel TFs: ATF6, E4F1, FOSB, HES1, HIF1A, NR1H3, RUNX2, SPIB.

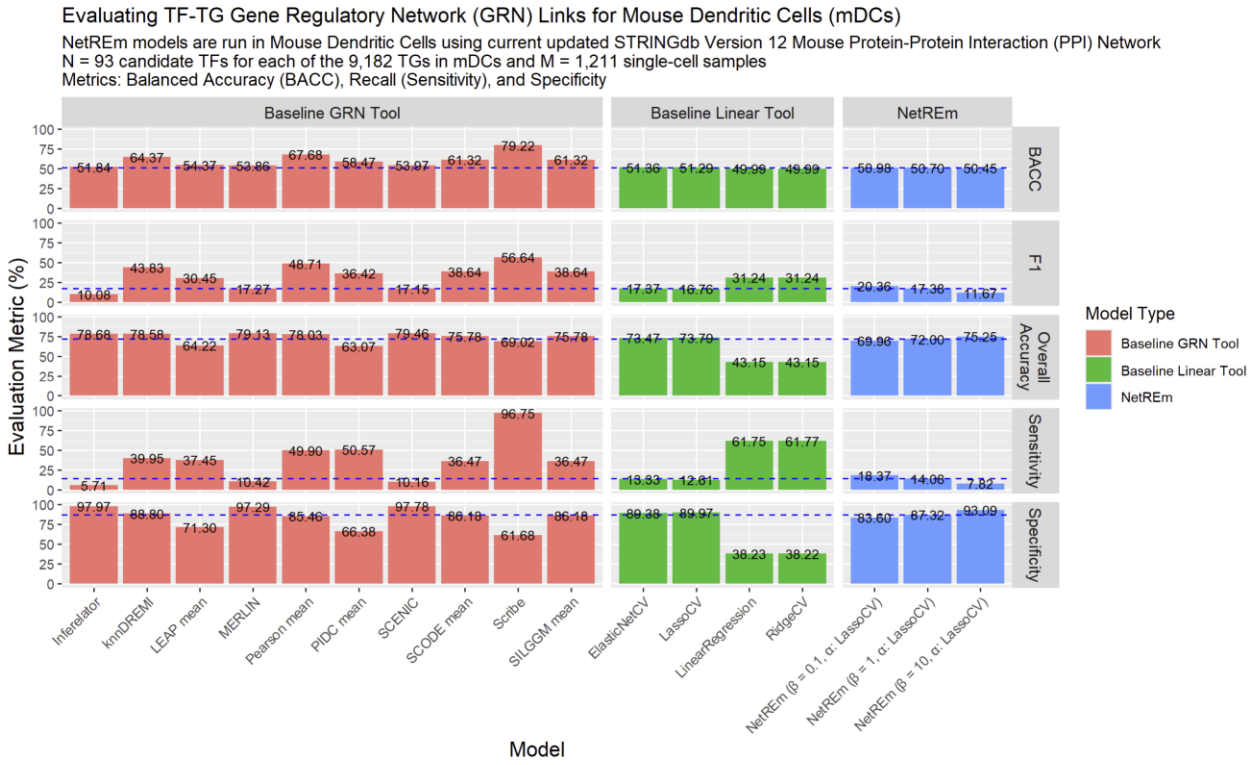


GO	Description	Log10(P)
WP236	Adipogenesis	-13.2
M279	PID RB 1PATHWAY	-11.0
WP5130	Th17 cell differentiation pathway	-10.8

Color	MCODE	GO	Description	Log10(P)
	MCODE_1	M279	PID RB 1PATHWAY	-7.4
	MCODE_1	WP236	Adipogenesis	-6.5
	MCODE_1	GO:0008285	negative regulation of cell population proliferation	-4.2

Figure B.7I) Novel TFs for *RNF167*: ARID5B, ATF7, MEF2C, NFE2L3, NR2C2, PATZ1, RFX1, STAT1, STAT3, TFAP4.
Novel TF TFAP4, is involved in Adipogenesis (creating adipocyte stem cells) and negative regulation of cell population proliferation with 3 substantiated TFs: E2F1, CEBPB, E2F4. Overall, NetREm can identify biologically-enriched PPI-subnetworks for TGs. Thus, NetREm has superior potential in revealing intricate regulatory networks for TGs.

Figure B.8 Analyzing NetREm's Predicted TF-TG Regulatory Links in Mouse Dendritic Cells (mDCs)



For TF-TF coordination, we look at network regularization hyperparameter $\beta = 1$ and 8 different NetREm models (varying sparsity hyperparameter α). For each metric, we have blue horizontal dashed lines that are for NetREm($\beta = 1$, LassoCV to determine α). Please note that metrics for the 14 baseline models (10 state-of-the-art cell-type GRN inference tools and 4 benchmark regression models (BRMs)) are available in **Table B.6**. As β increases from 0.1 to 10 (LassoCV for α), there are increases in: specificity and accuracy, decreases in F1 and sensitivity. For β = 10, there is a switch and NetREm has greater specificity and lower sensitivity relative to ElasticNet and Lasso. This illustrates trade-offs based on β tuning. Each of the 14 models has strengths and weaknesses, none superior to others. For β = 10, NetREm has higher sensitivity than Inferelator does and higher specificity than that of 11 other models (e.g. knnDREMI, SILGGM mean). For β = 1, NetREm has higher F1 scores than Inferelator, MERLIN, SCENIC's GRNBoost2. For β = 0.1, NetREm has higher accuracy than LEAP, PIDC, Scribe.

Figure B.9 Analyzing NetREm's ability to prioritize and predict future Protein-Protein Interaction (PPI) Links in Mouse Embryonic Stem Cells (mESCs).

Specifically, for mESCs, the fully-connected TF-TF PPIN in V11 contains 5,044 known and 13,871 artificial TF-TF edges among 195 candidate TFs; V12 contains 6,856 known and 12,059 artificial TF-TF edges; 25 singleton TFs in V11 PPIN have known PPIs in V12, indicating evolution from V11 to V12. 4 comparison boxplots of the 18,915 Cell-type TF-TF coordination scores ($\bar{B}$) in Mouse Embryonic Stem Cells (mESCs) based on the older STRING Protein-Protein Interaction (PPI) Network (PPIN) database version 11 (with averaged combined scores) as input to NetREm. The STRING database (db) considers direct as well as indirect interactions (i.e. functional associations) among the TFs. Outliers are removed from the boxplot. Input: v11 mouse PPN. Evaluation: updated v12 mouse PPIN. These 4 panels (**Figure B.9A-D**) analyze results of the following run: NetREm($\beta = 1, \alpha = 0.05$) was run on 19,225 TGs in mESCs (that each have single-cell gene expression measures $M = 1{,}080$ cells) using the same $N^* = 195$ candidate TFs for each TG; in cases where the TG is also a TF, that TF is removed and NetREm is run using 194 TGs. Missing pairwise TF-TF links are added to the input network (with default edge weight $\eta = 0.01$) to make the network fully-connected. We ran NetREm with the older mouse v11 PPIN as a proxy for older information (and held the newer network v12 as current, updated information to gauge NetREm's ability to detect future links).

- *Valid discoveries* are found in the newer, updated STRING v12, and were unknown in older v11 (input for NetREm); such discoveries show NetREm's ability to learn future PPI links of indirect/direct TF-TF interactions that may be unknown at the current time.
- *Unknown links* are those that have not been mentioned in both networks (i.e. STRING v11 and v12 do not mention them);
  - such links could nonetheless be *future discoveries* (False Negatives FNs)
    or
  - True Negatives (TNs: no meaningful direct/indirect interaction truly exists between them in biology).
- *V11 links* are found in STRING v11;
  - these links may still be in V12 networks: *Known (both)*
    or
  - may be *False Positives (FPs)* if they are no longer found in newer STRING v12 network: *Removed*).

Categories for TF-TF links based on mice STRINGdb PPI Networks (PPINs)		**Input** Version 11 (V11) PPIN (Outdated info)	
		Known in V11	Unknown in V11
Updated V12 PPIN (Future info)	Known in V12	Known (Both) *True Positive (TP)*	Valid Discovery *False Negative (FN)*
	Unknown in V12	Removed *False Positive (FP)*	Unknown

Figure B.9A-B focuses on the distribution of TF-TF coordination scores across these 4 potential groups: Removed, Unknown, Known (Both), and Valid Discovery. **Figure B.9C-D** collapses the Known (both) and Removed links into the V11 category and focuses on the distribution of TF-TF coordination scores across these 3 potential groups: Unknown, V11 links, and Valid Discovery. Please note: associated Statistical Tests for this analysis can be found in **Tables B.7-B.8.**

Figure B.9A) focuses on the magnitude of $\bar{B}$: $0 \leq |\bar{B}| \leq 100$.

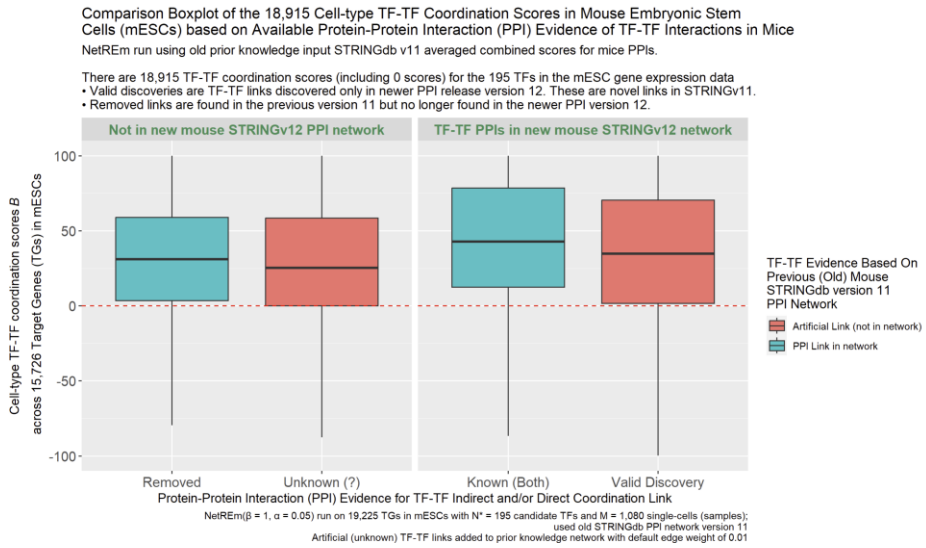


Figure B.9B) presents the original mESC TF-TF coordination scores: $-100 \leq \bar{B} \leq 100$.

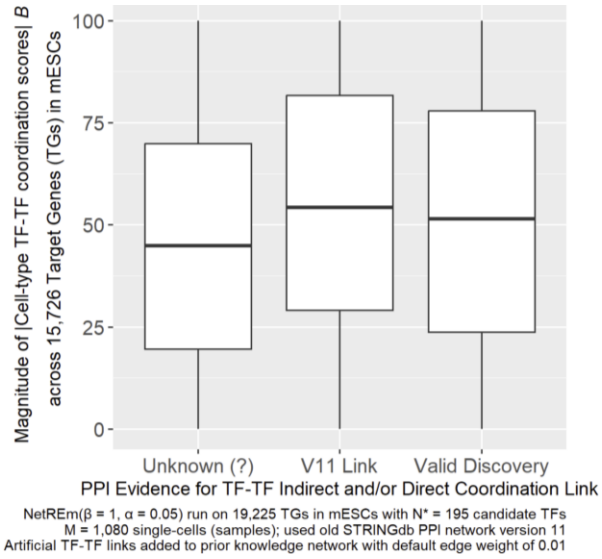


Figure B.9C) focuses on the magnitude: $0 \leq |\bar{B}| \leq 100$.

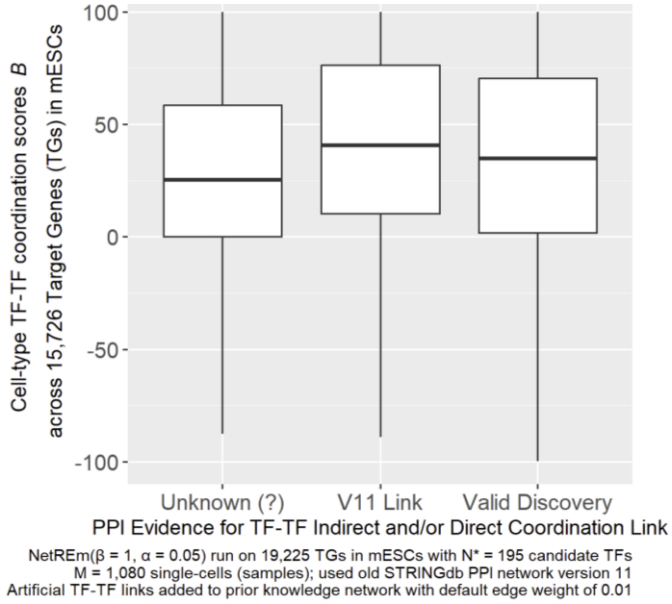


Figure B.9D) presents the original: $-100 \leq \bar{B} \leq 100$. The # of edges: 18,915 is based on: $(1/2)\ (N)(N-1)$.

Figure B.10 Analyzing NetREm's ability to prioritize and predict future PPIs in Mouse Dendritic Cells

Similar to **Figure B.9**, but we are focusing on mouse dendritic cells (mDCs), visualizing impacts for various sparsity prior hyperparameters α. Please see reference **Table B.7** (adapted for mESCs) and **Tables B.9-10**.

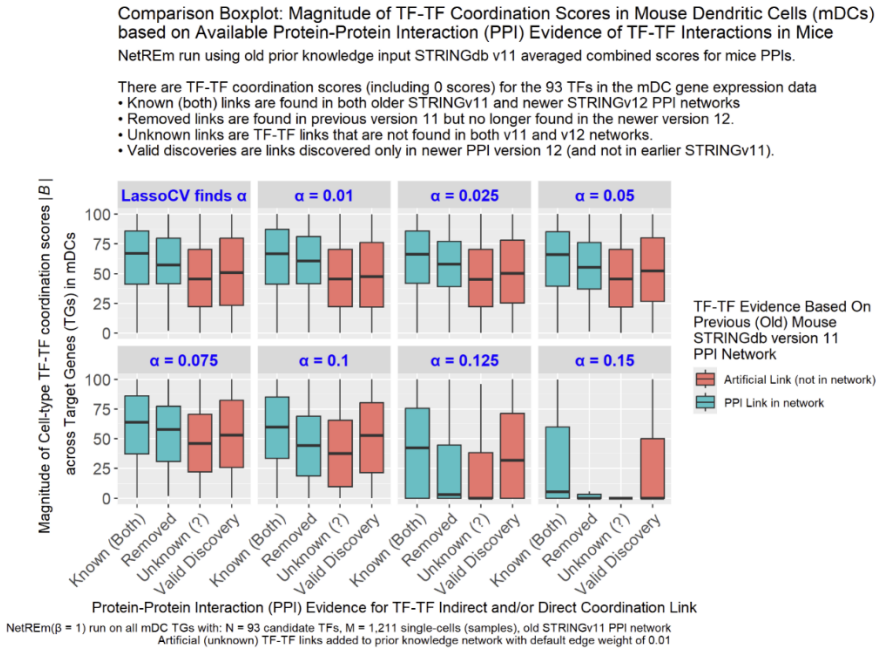


Figure B.10A) Comparison Boxplot: Magnitude of TF-TF coordination $0 \leq |\bar{B}| \leq 100$ scores in mDCs for the 4 categories of TF-TF Links: Known (Both), Removed, Unknown, and Valid Discovery. The results of NetREm runs for different values of the sparsity prior hyperparameter α.

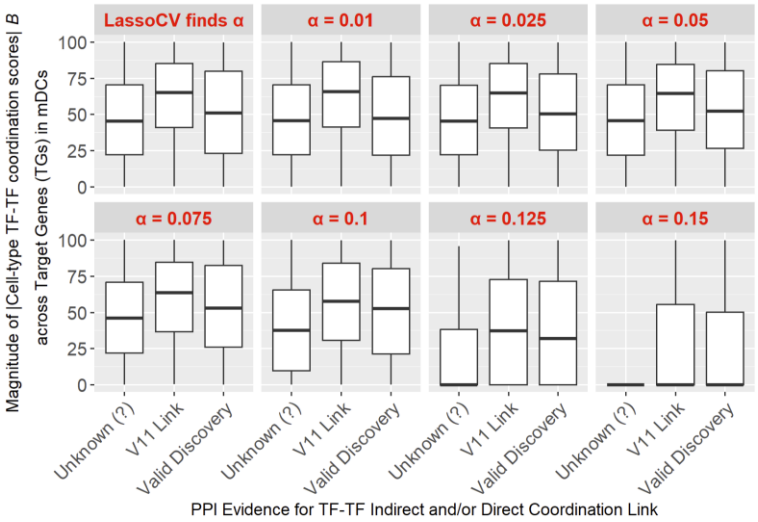


Figure B.10B) Comparison Boxplot: Magnitude of TF-TF coordination $0 \leq |\bar{B}| \leq 100$ scores in mDCs for 3 categories of TF-TF links: Unknown, V11 Link, Valid Discovery.

The results of NetREm runs for different values of the sparsity prior hyperparameter α. The V11 Link category includes Known (Both) and Removed TF-TF links.

Figure B.11 NetREm TF-TF Coordination Performance for 9 different Human Immune Cell Types: PBMCs: Peripheral Blood Mononuclear Cells.

For PBMCs, we intersect the 13,714 TGs with (Lambert et al. 2018) to uncover 1,029 candidate TFs, which we say is $\mathcal{N}$. Then, for most TGs, $N = \mathcal{N} = 1{,}029$ and if the TG is also 1 of the 1,029 TFs then $N = \mathcal{N} - 1 = 1{,}028$ candidate TFs. We run NetREm using the older version 11 STRINGdb PPIN.

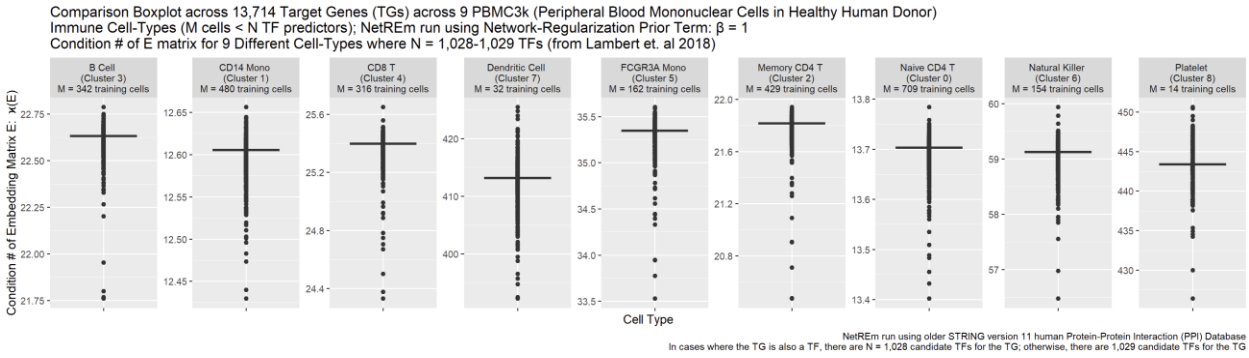


Figure B.11A) Comparison boxplot across 13,714 target genes across 9 PBMC Immune cell-types *(*M cells < N = 1,029 candidate TFs).

NetREm is run using $\beta = 1$ and LassoCV to optimize α. We can see the condition #s over the 13,714 TGs for each of these 9 types. Please note the max condition # of the E matrix is: $\max(\kappa(E))$ = Naive CD4 T (Cluster 0): 13.8, CD14 Mono (Cluster 1): 12.7, Memory CD4 T (Cluster 2): 21.9, B (Cluster 3): 22.8, CD8 T (Cluster 4): 25.6, FCGR3A Mono (Cluster 5): 35.6, Natural Killer (Cluster 6): 59.9, Dendritic Cell (Cluster 7): 425, Platelet (Cluster 8): 451. Since $\max(\kappa(E)) \ll 1e6$, all cells are very well-conditioned and therefore are numerically stable. In particular, the $M = 14$ versus N = 1,029 case for Cluster 8 (Platelets) illustrates this, given that $M \ll N$.

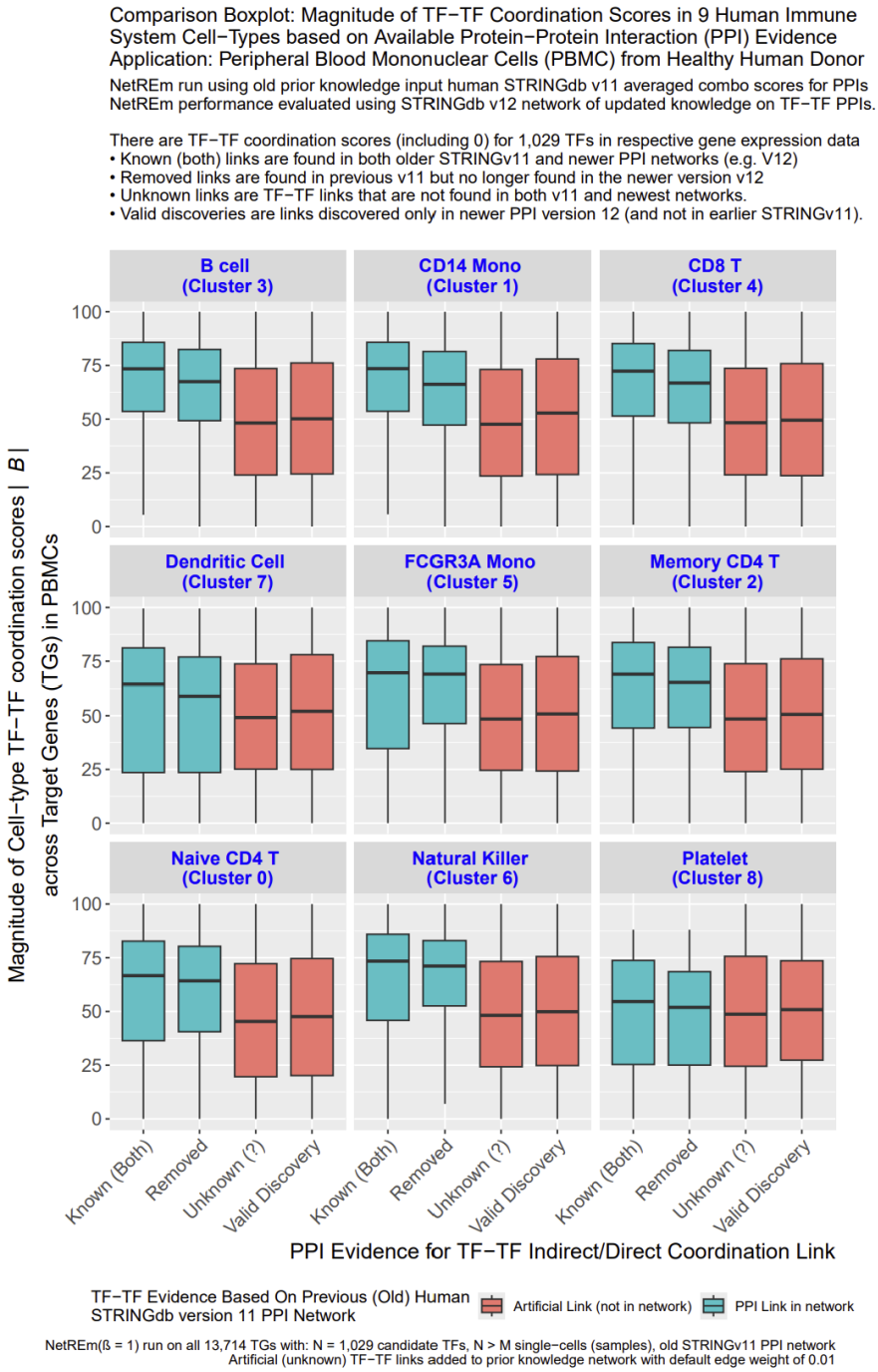


Figure B.11B) Comparison Boxplot: Magnitude of TF-TF coordination scores in 9 human immune PBMCs based on available PPI evidence.

NetREm run using old V11 PPIN and evaluated using the updated human PPI network (including STRINGv12 network and other resources) on TF-TF PPIs. There are TF-TF coordination scores including 0 for 1,029 TFs in respective gene expression data. The known (both) links are found in the older V11 and newer PPI networks. Removed links are found in V11 but no longer in the newest network. Unknown links are those that are not found in both networks. Valid discoveries (VDs) are not in V11 but uncovered in newest network. The y-axis is the magnitude of cell-type TF-TF coordination scores $|\bar{B}|$.

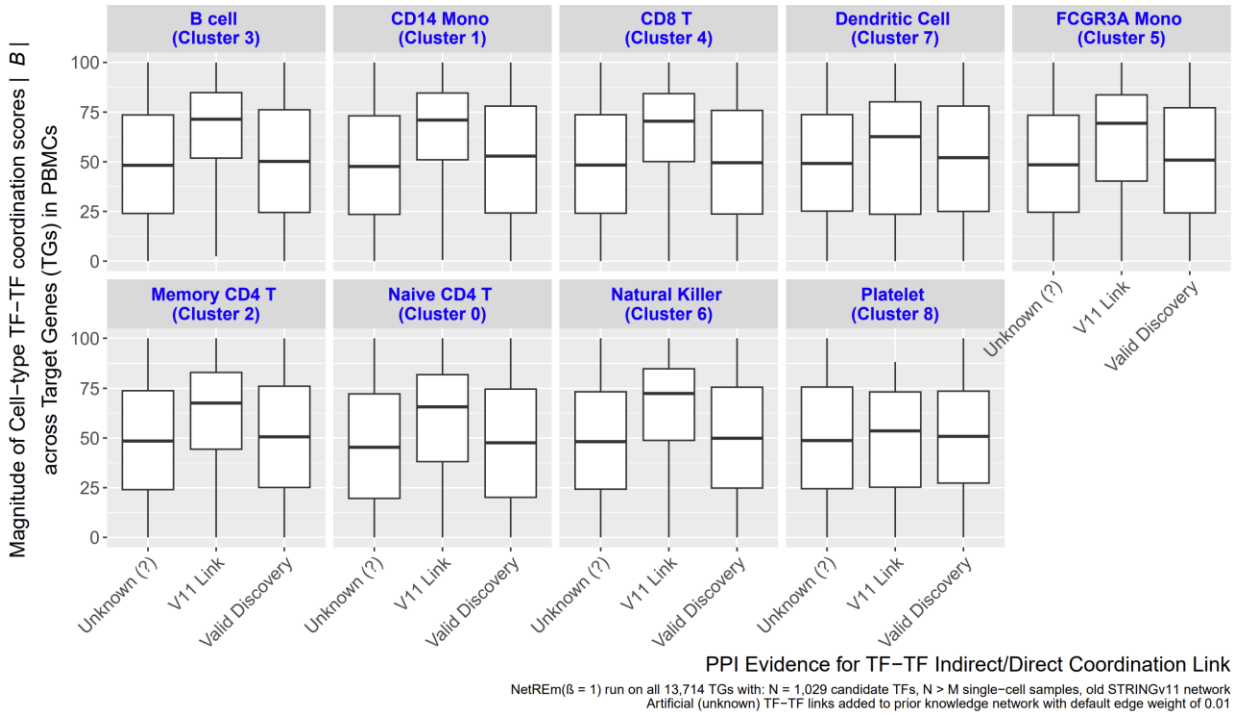


Figure B.11C) Same as **Figure B.11B**, except we collapse 2 TF-TF subgroups.

Essentially, the V11 links could be known links (in V11 and newest network) or they could be removed. The other 2 (Unknown and Valid discoveries) are still focusing on links not in the older V11 network: still not in the newest network (if they are unknown) or have just been discovered in the newest network (if they are valid discoveries).

Figure B.12 Benchmark: Compare predicted TF-TF Links for NetREm versus RTNduals

We retrieve a certain # of final TF-TF links by RTNduals(Chagas et al. 2019), say k. Then, we go to NetREm's cell-type TF-TF coordination network and extract the top k links in terms of their magnitude of cell-type TF-TF coordination scores $|\bar{B}|$. So, we ultimately compare NetREm's top k links with RTNduals' final k links. We gather our comprehensive input PPIN for humans (Based on STRINGdb V12 and additional sources). Then, we add in novel links found by (Göös et. al, 2022) on strong TF-TF physical PPIs (based on high SAINT scores). We use this *newest PPIN* to evaluate NetREm versus RTNduals. TF-TF coordination can be indirect or direct. Here, we use an older PPIN in humans (e.g. STRINGdb v11) and try to determine how many links are novel discoveries (not in V11 PPIN but found in *newest PPIN*). We also see how many links are in both networks (V11 and *newest PPIN*). Then, we see how many of these links are poor results: unfortunately False Positives FPs (in V11 PPIN but removed in *newest PPIN*) and/or still unknown. In this way, we account for TF-TF coordination links that correspond to direct/indirect PPIs.

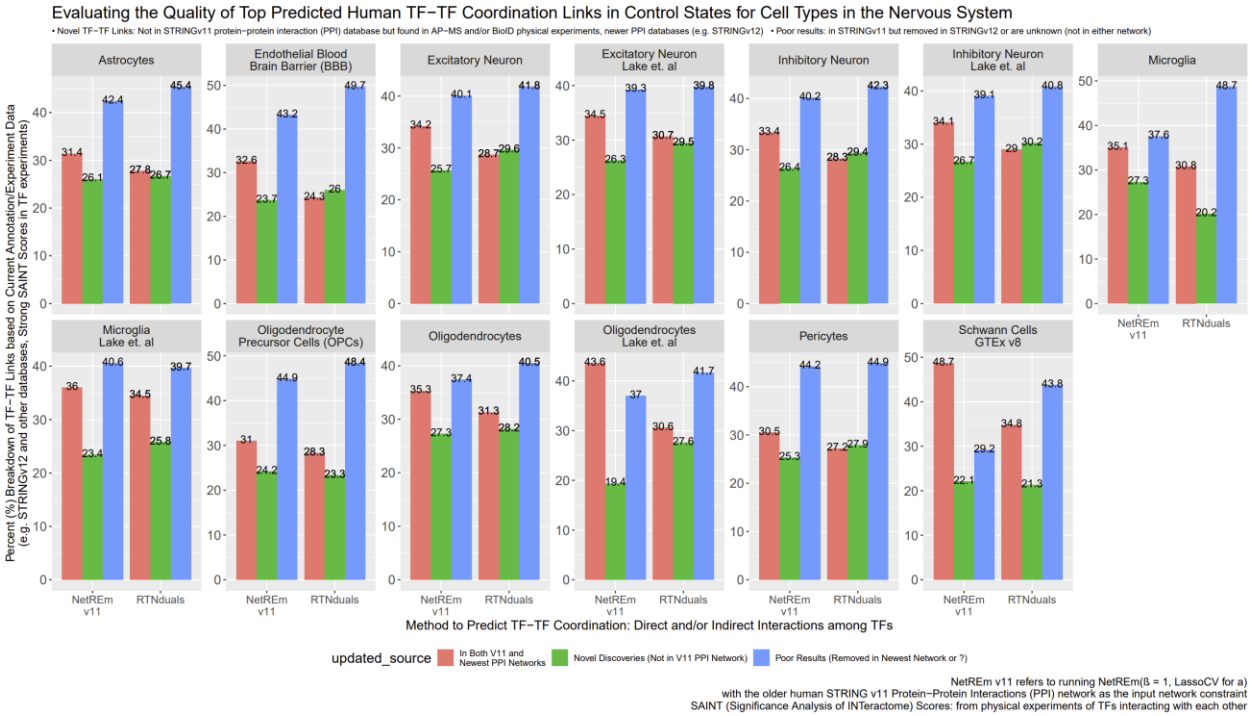


Figure B.12A) We compare results for NetREm with V11 input PPIN across 13 human contexts (i.e. different cell-types).

Except for Microglia (Mic) (Lake et al. 2018), the other 12 contexts show NetREm has fewer poor results than RTNduals does. NetREm uses known prior info of TF-TF PPIs (though some PPIs may be removed in future releases), so it naturally performs better in prioritizing actual links. For instance, in Control Mic, NetREm has 37.6% of TF-TF links that are poor results, 35.1% of top k links are in both PPINs, remaining 27.3% of links are valid novel discoveries (not in V11, in updated PPIN); in comparison, 48.7% of Control Mic links predicted by RTNduals are poor. Similarly, for GTEx SCs 29.2% of NetREm's links are poor, compared to 43.8% for RTNduals.

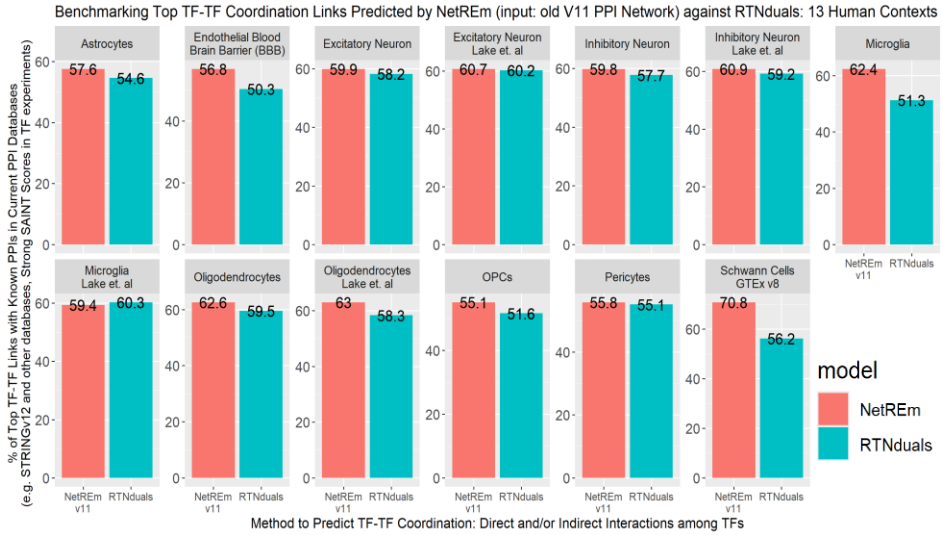


Figure B.12B) Summarizing **Fig. B.12A**: comparing the good results between RTNduals and NetREm. We note that in 12 of 13 contexts, NetREm uncovers a greater % of TF-TF links that are good (biologically meaningful, known TF-TF PPI links) results.

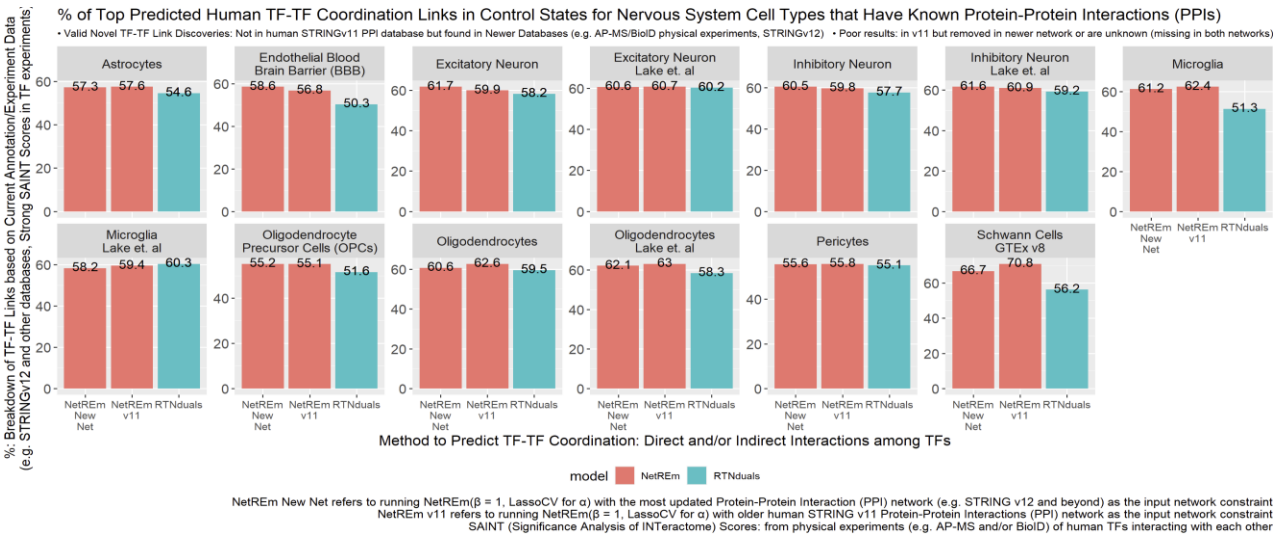
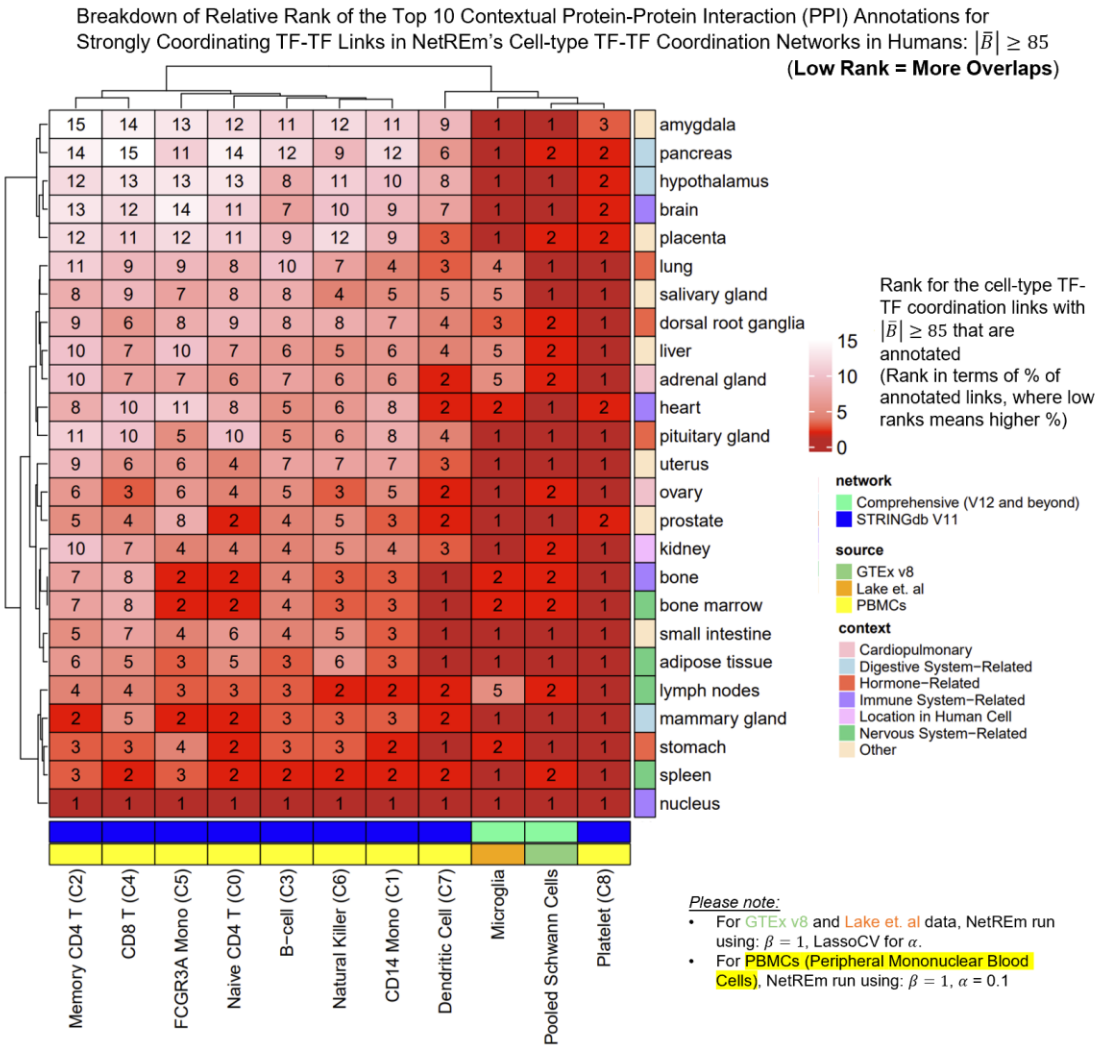


Figure B.12C) In addition, we run NetREm on newest PPIN and compare the 3 approaches (RTNduals, NetREm with V11 PPIN, NetREm with newest PPIN) across these 13 datasets.

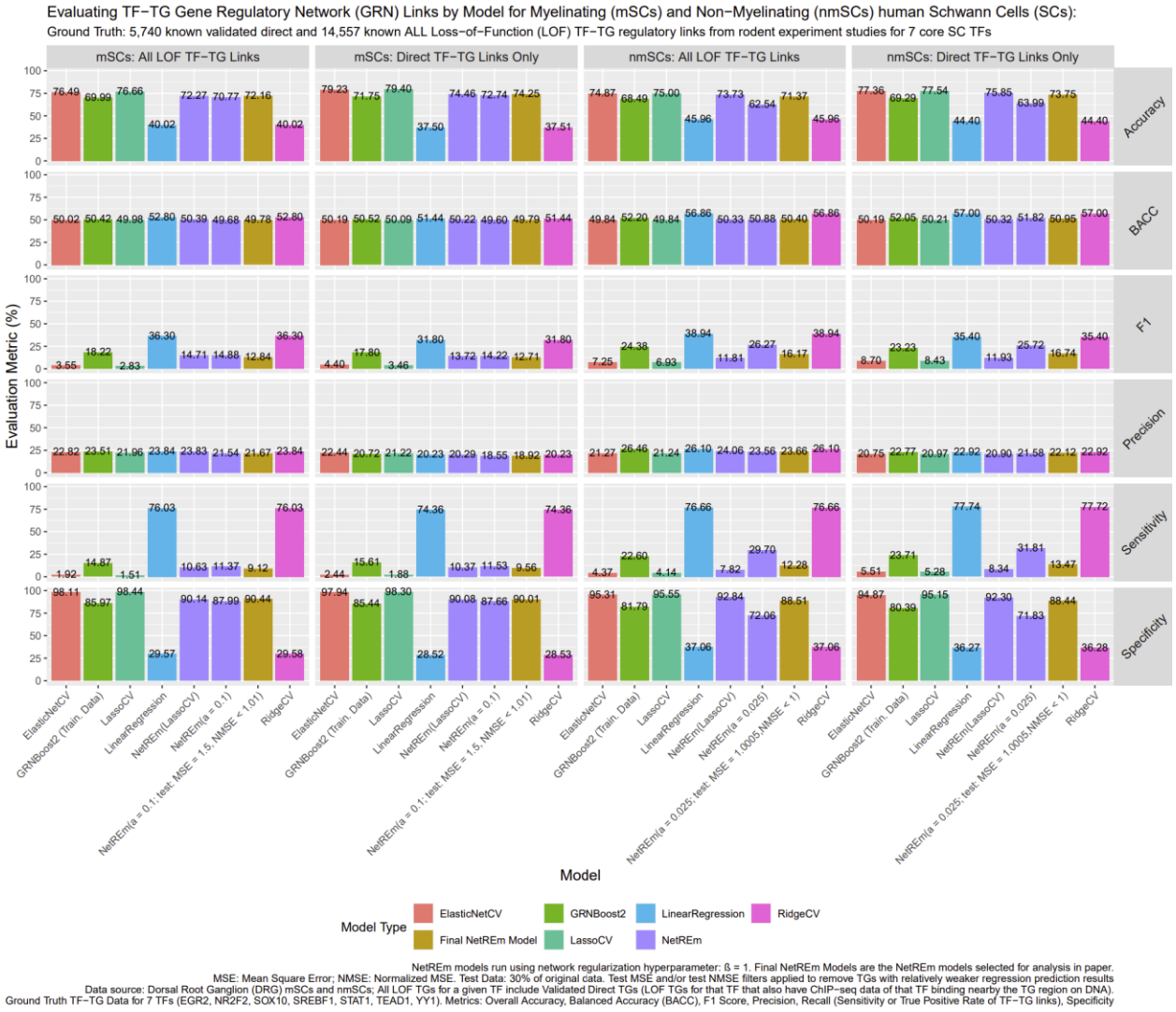
Overall, both NetREm approaches always learn a greater % of annotated PPI links that are in the newest PPIN (i.e. they have fewer poor results than RTNduals does). By using known PPIs (although some may be FPs), NetREm better prioritizes actual links. Thus, NetREm based on an older PPIN can predict and prioritize TF-TF links in a newer PPIN, so NetREm's top novel TF-TF coordination scores may predict potentially undiscovered links (i.e. identified in future releases). NetREm, even on older V11 PPIN, more effectively identifies and prioritizes TF-TF interactions validated in V12 PPINs and beyond.

Figure B.13 Contextual PPI Database Annotations for Top TF-TF Coordination Links



We note that the TF-TF coordination networks can prioritize cell-specific TF-TF coordination. Based on our revisions to our method, we now have cell-type TF-TF coordination scores that can fall between -100 and 100 (i.e. $|\bar{B}| \leq 100$). To show this, we do the following analysis for several cell-types in humans. For the given cell-type, we retrieve cell-type TF-TF links that are relatively strong $|\bar{B}| \geq 85$. We overlay those with annotated TF-TF links in the Contextual PPI database(Kotlyar et. al 2022), computing the % of top TF-TF links enriched for each of the 240 different annotations. Then, we rank the results, where low ranks represent a greater %. We note all of the top TF-TF links are predicted to localize in the nucleus, consistent with their functions as TFs. For instance, in Microglia and pooled Schwann cells (SCs), we find our strong TF-TF coordination links are more typically more enriched for nervous system-related terms (e.g. dorsal root ganglia, brain, hypothalamus, amygdala) than 9 immune sub-populations in PBMCs (immune-related cells) are. These PBMC cell-types tend to have strong TF-TF coordination links that are most strongly enriched for immune-related terms (e.g. spleen, bone marrow, lymph nodes, adipose tissue). We note how different PBMC sub-populations have more enrichment (lower rank values) for certain immune-related terms than others, possibly suggesting their localization and functions in the immune system. Given that we use Pooled GTEx SCs for this analysis (which are gathered from 5 tissues: esophagus mucosa/muscularis, heart, prostate, skeletal muscle), our top TF-TF coordination links for SCs tend to be highly enriched for relevant tissues (e.g. heart, prostate).

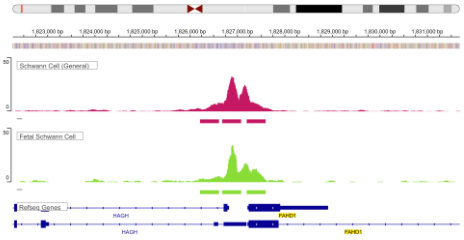
Figure B.14 NetREm TF-TG regulatory link performance metrics for 7 core Schwann cell (SC) Transcription Factors (TFs)



This analysis is done for only the 7 TFs that we have background, experimental validation data for: EGR2, NR2F2, SOX10, SREVF1, STAT1, TEAD1, YY1. Please note that for the Loss-of-function (LOF) evaluation, there are 64,765 total possible Loss-of-Function TF-TG regulatory links in SCs. 14,557 are total positive links that appear in the ground truth GRN.

Figure B.15 Humans: Open Chromatin in human SCs and SOX10 predicted binding regions for 4 novel SOX10-predicted TGs in mSCs

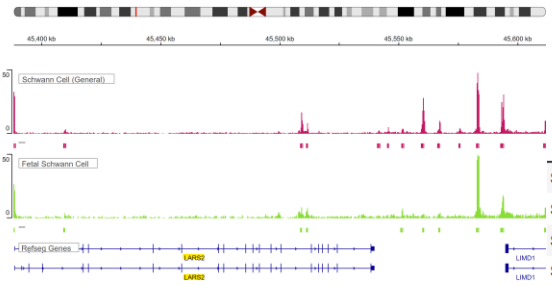
Binding peaks of open chromatin in humans (hg38 reference genome) based on (Zhang et al. 2021) for the corresponding 4 novel final TGs for SOX10 in SCs shown in **Figure 3.3D** (rat validation data). None of them are LOF or direct TGs for SOX10. Left: epigenome tracks of accessible chromatin in humans. Top (scarlet): adult general SCs, bottom (green): fetal SCs. Right: table of predicted SOX10 binding regions for each of these TGs in the input prior GRNs in adult mSCs derived from multiomics data (initial TF-RE-TG links).



predicted SOX10 binding regions for *FAHD1*

TF	TG	chromosome	start	end	region	genome
SOX10	FAHD1	chr16	1826388	1827564	promoter	hg38

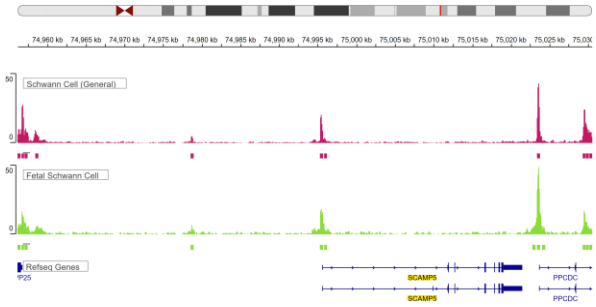
Figure B.15A) SOX10 may bind to a promoter region for *FAHD1*.



predicted SOX10 binding regions for *LARS2*

TF	TG	chromosome	start	end	region	genome
SOX10	LARS2	chr3	45388244	45388873	Distal (Enhancer)	hg38
SOX10	LARS2	chr3	45511534	45511853	Distal (Enhancer)	hg38
SOX10	LARS2	chr3	45582929	45583865	Distal (Enhancer)	hg38
SOX10	LARS2	chr3	45611555	45611991	Distal (Enhancer)	hg38

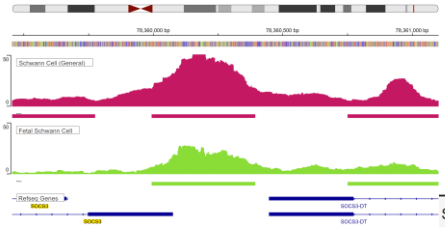
Figure B.15B) SOX10 may bind to various potential enhancers for *LARS2*.



predicted SOX10 binding regions for *SCAMP5*

TF	TG	chromosome	start	end	region	genome
SOX10	SCAMP5	chr15	74956149	74957552	Distal (Enhancer)	hg38
SOX10	SCAMP5	chr15	75029296	75030505	Distal (Enhancer)	hg38

Figure B.15C) SOX10 binds to various potential enhancer regions for *SCAMP5*.



predicted SOX10 binding regions for *SOCS3*

TF	TG	chromosome	start	end	region	genome
SOX10	SOCS3	chr17	78359457	78361101	promoter	hg38

Figure B.15D) SOX10 may bind to a potential promoter for *SOCS3*.

Figure B.16 Complex barplot showing the top 14 Transcription Factors (TFs) with the highest # of Tibial Nerve eQTL-validated TGs overall across myelinating (mSCs) and non-myelinating (nmSCs) SCs

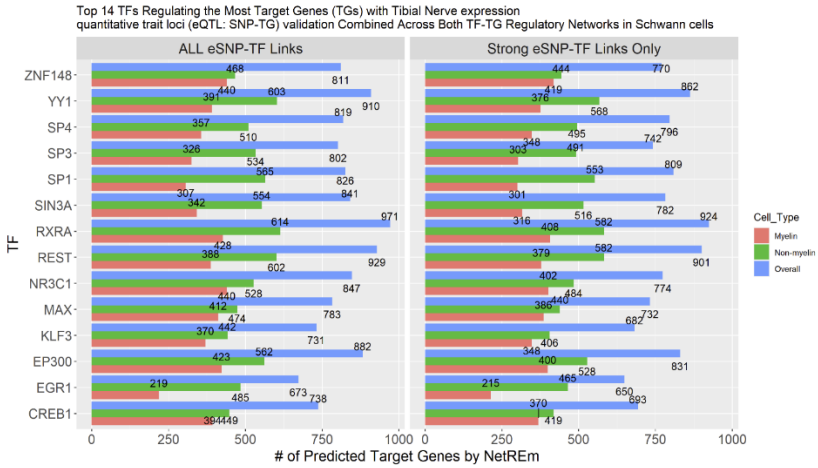


Figure B.16A) This comparative plot focuses on the top 14 TFs that altogether regulate the greatest # of unique TGs that have eQTL validation in SCs.

To this end, we pool NetREm's results from mSCs and nmSCs. This plot shows the 14 TFs with the highest # of eQTL-validated TGs (eTGs) in both mSC and nmSC TF-TG regulatory networks (complementary GRNs). Left panel: TFs with this validation based on strong and/or weak SNP-TF effects predicted by motifbreakR tool; right panel: TFs with this validation based on only the strong predicted SNP-TF effects. YY1 regulates 910 TGs overall (weak and/or strong support by eQTLs), of which 862 TGs overall have strong eQTL support.

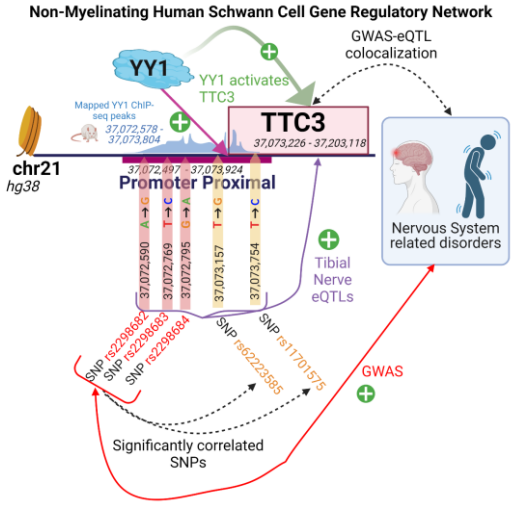


Figure B.16B) Examples of eQTL SNPs linked to changes in expression of Target Gene (TG) *TTC3* in nmSCs by altering YY1's ability to bind to TF Binding Sites to regulate *TTC3* in humans.

Here, we present an example of eTG *TTC3* that is regulated by YY1 (one of top TFs regulating TGs with eQTL support) in nmSCs. YY1 binds to a promoter proximal region of *TTC3* to activate *TTC3* ($c^* > 0$). We find eQTL SNPs significantly correlated with each other that also link to increased *TTC3* expression; 3 of these are risk SNPs for nervous system (NS)-related disorders. *TTC3* has strong GWAS-eQTL colocalization with NS-related disorders. We find ChIP-seq peaks support YY1 binding to this orthologous region in rodents.

Figure B.17 TF-TF Interactions and TF-TG Regulatory Networks in Human Schwann cells

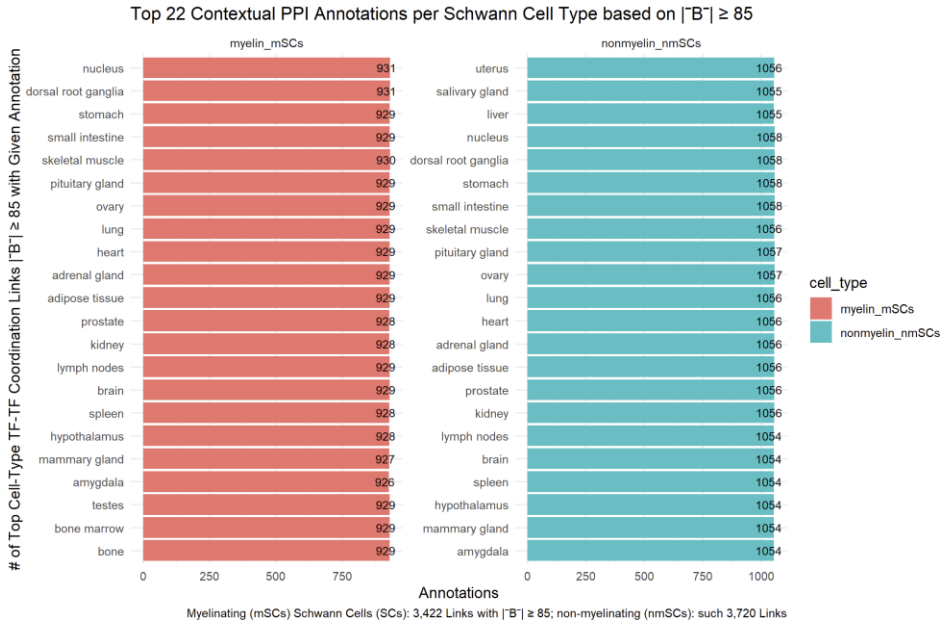


Figure B.17A) Top 22 annotations (in terms of overlap) for strong cell-type TF-TF coordination scores (e.g. $|\bar{B}| \geq 85$*)* based on Contextual PPI (Kotlyar et. al 2022) database (CPPID). Here, 243 annotations were tested.

PPI Edge	Peripheral nervous system neoplasm	Other nervous system diseases or cancers	DRG / brain
	?	?	?
	?	?	✓
	?	✓	✓
	✓	✓	✓

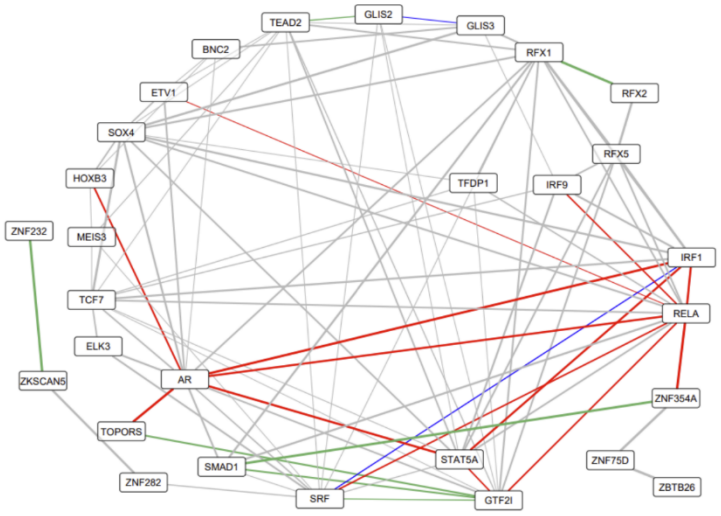


Figure B.17B) TF-TF Interaction Network of 29 of 31 non-myelinating SCs: nmSC-specific TFs, with edge weights representing averaged B-matrix coordination scores across the 5,164 TGs in final nmSC NetREm TF-TG Regulatory Network.

There are 93 TF-TF links here. Since TF-TF coordination links shown are subset to those known in input PPI network (PPIN), we may refer to these links as TF-TF interactions (i.e. those with evidence of known PPIs). Higher cell-type coordination scores ($\bar{B}$) indicate stronger predicted TF-TF interactions. We annotate TF-TF interaction edges based on CPPID. Coordination scores predicted for 8,943 mSC TGs, all 5,207 nmSC TGs.

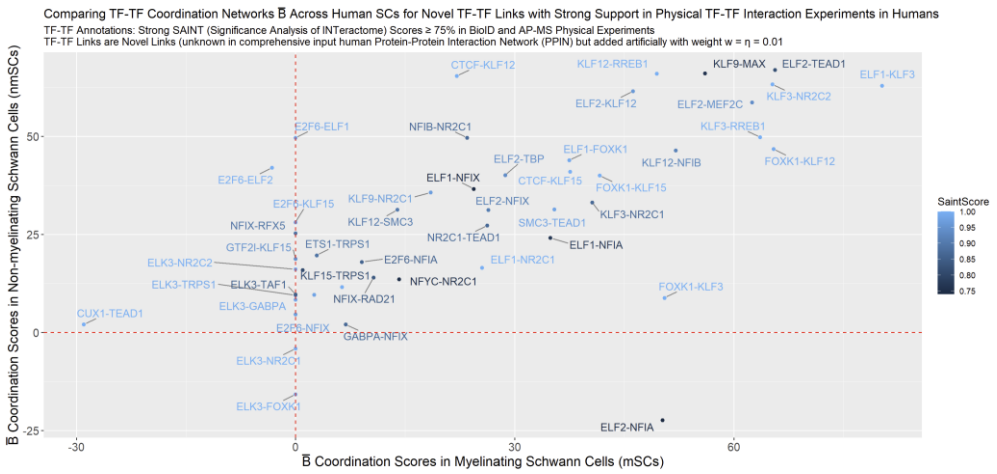


Figure B.17C) Comparative scatterplot of 48 novel TF-TF coordination links (unknown in comprehensive input human PPIN but added artificially to make input PPIN fully-connected, with $w = \eta = 0.01$) that are found in mSCs (37) and/or nmSCs (all 48).

These links comprise 30 different TFs. Here, these novel TF_i-TF_j links are discovered in (Göös et al. 2022) recent study on direct, physical human TF interaction networks where these TFs are predicted to physically bind in AP-MS and/or BioID experiments with very strong SAINT scores of interactions. Please note that other novel TF-TF links could correspond to indirect TF-TF interactions that are not yet known (since we used the most comprehensive version). Overall, this shows NetREm's novel TF-TF coordination links may be biologically meaningful discoveries and also shows differences between the cell-type TF-TF coordination scores $\bar{B}$ between human mSCs and nmSCs.

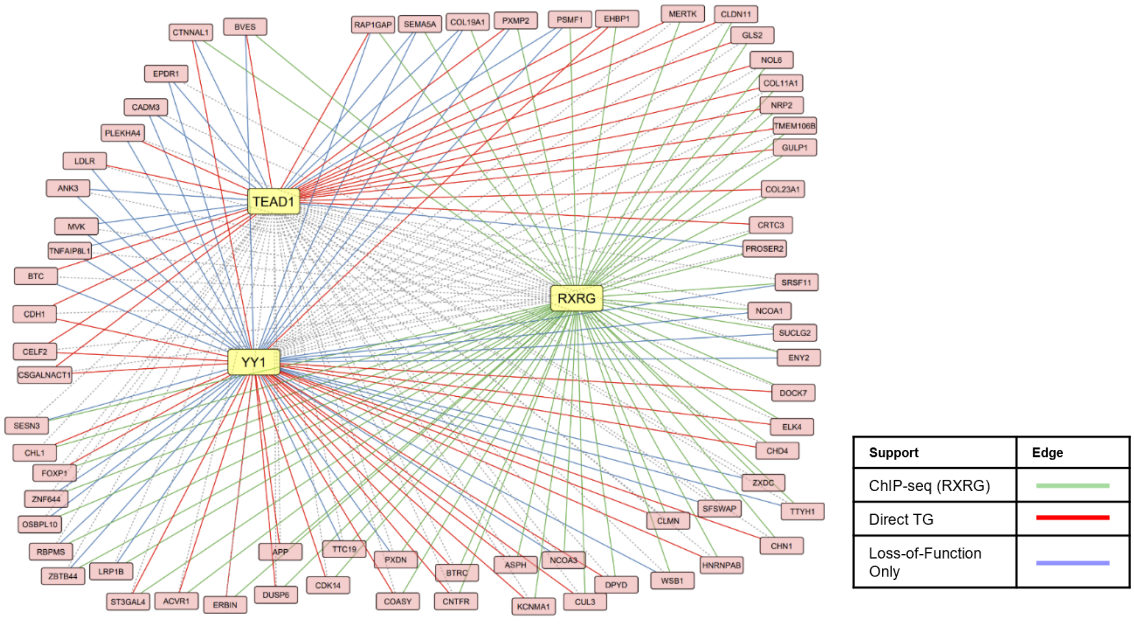
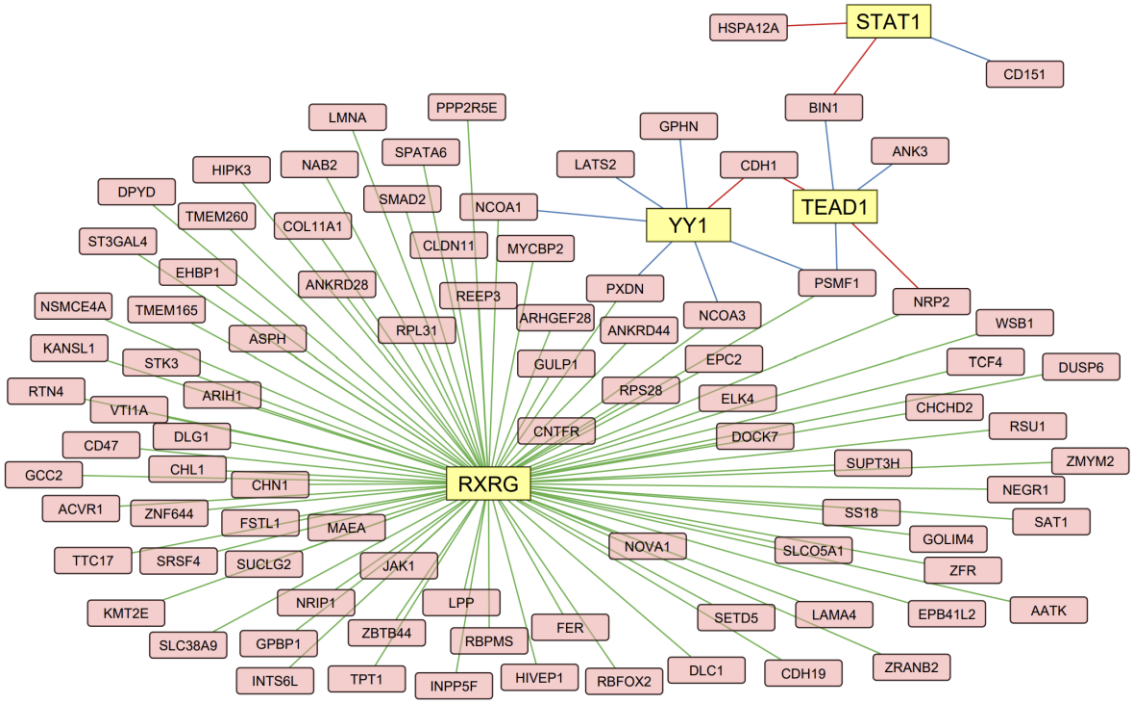


Figure B.17D) 68 of 366 common TGs regulated by 3 Core SC TFs (YY1, RXRG, TEAD1) in nmSCs.

We only have ChIP-seq data for RXRG and do not have knowledge of validated Direct Target Genes (TGs) or Loss-of-Function (LOF) TGs for RXRG. For visual simplicity, we truncated the list of 366 common TGs to those that have validation data in at least 2 of these 3 TFs. That is, out of the 5 possible sources, it is found in

at least 2 sources for at least 2 of the 3 TFs: ChIP-seq peak for RXRG, LOF only or Direct Valid for TEAD1, LOF only or Direct Valid for YY1. In the process genes like *CDH19* (RXRG ChIP-seq TG) and *NCAM1* (YY1 LOF only TG) are pruned out. Edges in red and/or blue correspond therefore to YY1 and/or TEAD1 TGs. The red edges are those with support that the TG is a validated direct TG for that TF (most confident TGs for the TF: support as LOF TG for the TF and ChIP-seq binding support) while those in blue are only LOF TGs for the TF (do not have strong ChIP-seq binding support). For instance, *EHBP1* is a Direct TG for YY1 and TEAD1 with ChIP-seq support for RXRG and NetREm predicts these 3 TFs regulate *EHBP1* in nmSCs.



Support	Edge
ChIP-seq (RXRG)	
Direct TG	
Loss-of-Function Only	

Figure B.17E) Common Target Genes (TGs) regulated by 4 Core Schwann Cell (SC) TFs (STAT1, YY1, RXRG, TEAD1) in nmSCs.

All 4 TFs co-regulate 174 TGs in nmSCs. Here, this network visualization is filtered to only show the edges that have experimental validation support. RXRG only has ChIP-seq data (and no loss-of-function (LOF) information) available.

Figure B.18 Analyzing NetREm TF-TF predictions and validations for 8 core Schwann cell (SC) Transcription Factors (TFs)

In this analysis, we have experimental data for 8 core SC TFs: EGR2, NR2F2, RXRG, SOX10, SREBF1, STAT1, TEAD1, and YY1. For some TFs we have more experimental data than we do for others. EGR2 is not present in non-myelinating SCs (nmSCs). Thus, figure panels **Fig. B.18A-F** do not have EGR2 for the columns; these heatmaps can be interpreted where the top diagonal (top right) refers to myelinating Schwann cells (mSCs) and the bottom diagonal (low left) refers to non-myelinating SCs (nmSCs). We use rat data in SC lines PNS (Peripheral Nervous System) and/or S16 to help validate our findings (**Fig. B.18G-H**).

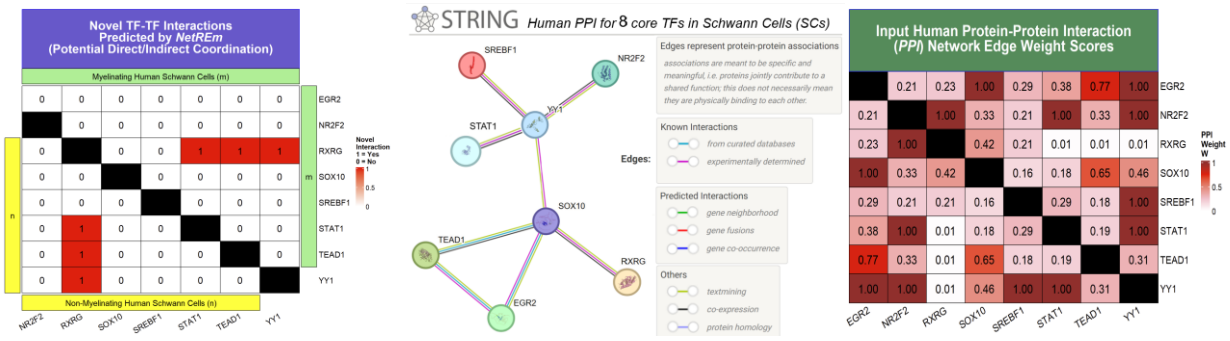


Figure B.18A) - NetREm predicts 3 novel TF-TF interactions (not found in the input Protein-Protein Interaction (PPI) network), namely: RXRG-STAT1, RXRG-TEAD1, and RXRG-YY1.

Middle panel (run from string-db.org session) illustrates how current PPI databases like STRING v12 (one of the resources for our input PPI) do not have any edges between RXRG-YY1, RXRG-STAT1, and RXRG-TEAD1. Given that the input PPI is comprehensive, we uncovered PPI edges among remaining TFs. Right panel: current input PPI edge weights provided (using STRINGdb and other sources) with default weight $w = \eta = 0.01$ for missing edges.

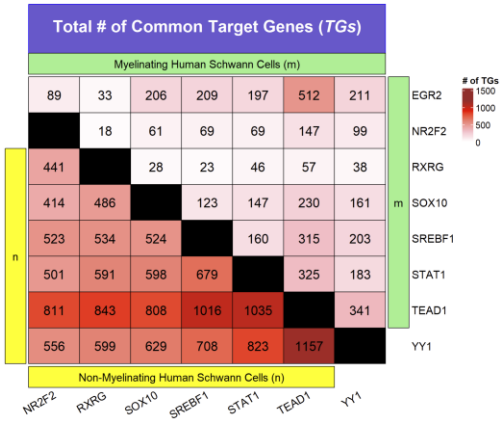


Figure B.18B) Common # of Target Genes (TGs) predicted to be co-regulated by the given pair of TFs. Coordinating TFs tend to share a significantly larger # of TGs than expected by chance (Wang et al. 2009).

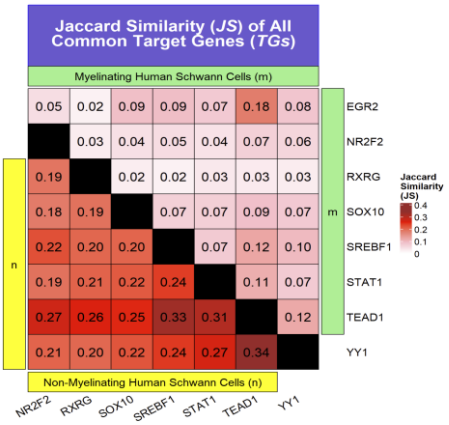


Figure B.18C) Jaccard Similarity (JS) of the shared TGs, where each pairwise JS is the # of shared TGs divided by the total # of TGs predicted across both TFs.

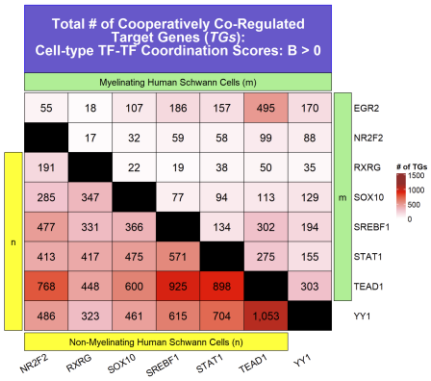


Figure B.18D) # of TGs that are predicted to be cooperatively co-regulated (i.e. $\bar{B} > 0$*)* by the given pair of TFs (i.e. set of TGs they have similar signs of coefficients c^* for). That is, for each $TF_i - TF_j$ pair we count the # of co-regulated TGs where $c_i^* c_j^* > 0$.

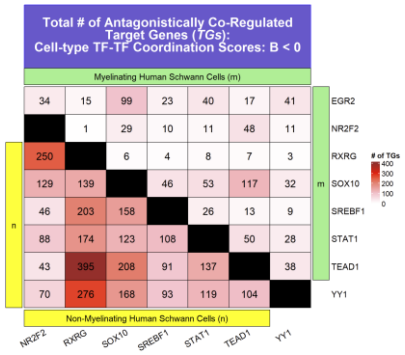


Figure B.18E) # of TGs that are predicted to be antagonistically co-regulated (i.e. $\bar{B} < 0$) by the given pair of TFs (i.e. set of TGs they have opposite signs of c^* for). That is, for each $TF_i - TF_j$ pair we count the # of co-regulated TGs where $c_i^* c_j^* < 0$.

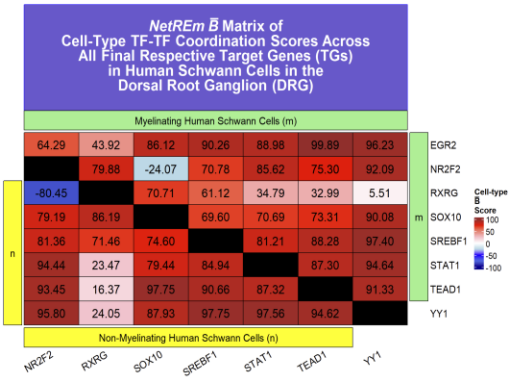


Figure B.18F) Final cell-type TF-TF coordination scores predicted by NetREm for these core SC TFs. This is based on averaging the TF-TF coordination scores for 8,943 TGs in mSCs, 5,207 TGs in nmSCs. TFs with non-zero coordination scores for a TG are predicted = as final model TFs for the TG.

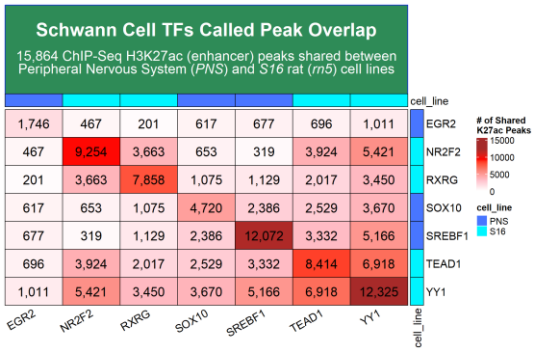


Figure B.18G) SC TFs called peak overlap based on 15,864 ChIP-seq enhancer peaks (H3K27ac) shared between PNS and S16 rat cell lines. This is the only symmetric heatmap in **Figure B.18**.

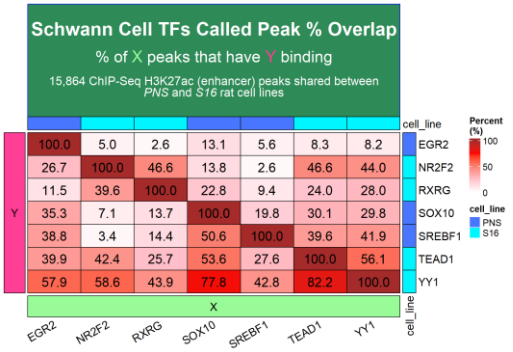
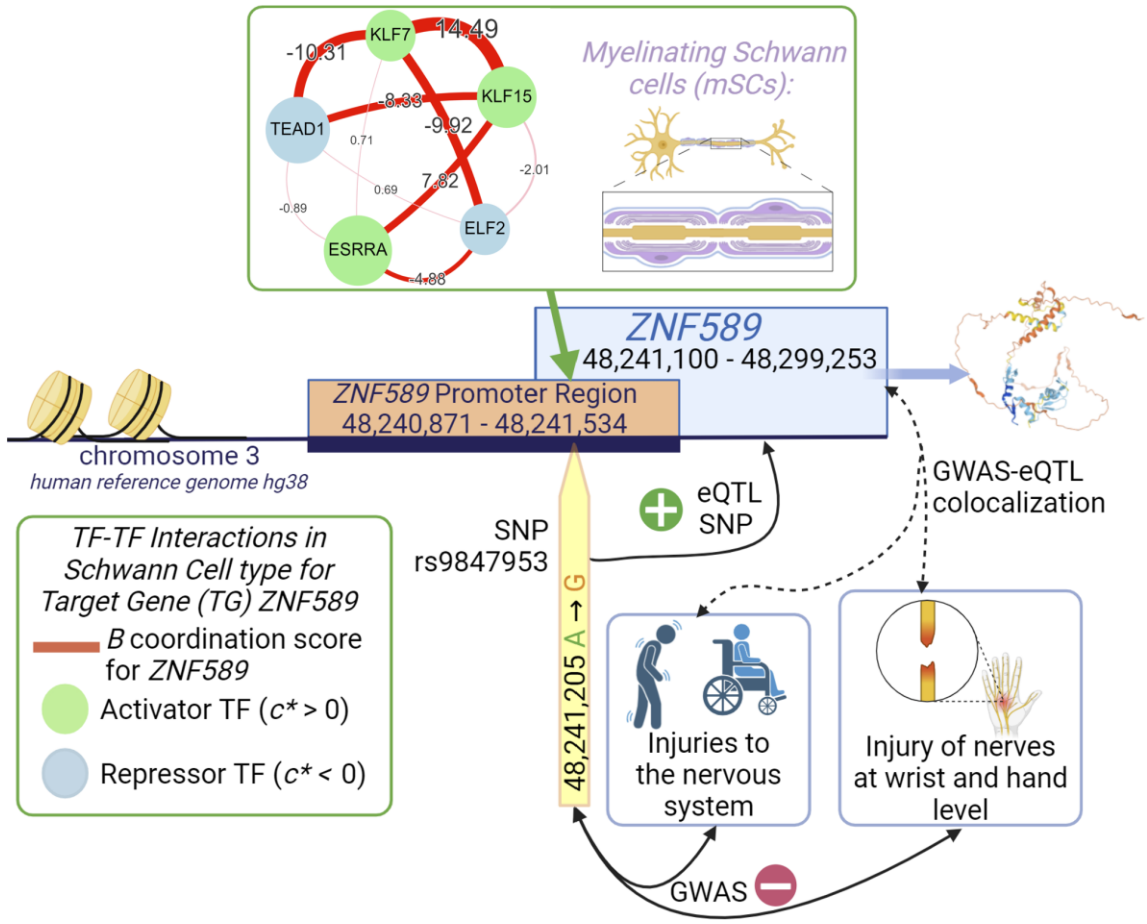


Figure B.18H) Schwann Cell TFs Called Peak % (Percentage) Overlap.

Based on **Figure B.18G** but presents binding %. Columns refer to X axis, and rows refer to the Y axis. The plot is interpreted by going to the X axis and selecting a TF. Then, we go up the row to the TF of interest (our Y). So, the 1st column (X = 1: EGR2), 2nd row from the top (Y = 2: NR2F2) refers to the % of shared SC peaks for EGR2 that have also have strong binding for NR2F2: ~26.7%. Conversely, the 2nd column (X = 2: NR2F2), 1st row from the top (Y = 1: EGR2) is 5% so ~5% of peaks for NR2F2 have strong binding for EGR2 as well.

Figure B.19 TF-TF Coordination and TF-Regulatory_Element-TG Regulatory Networks with eQTL validation in myelinating Schwann cells (mSCs) for a disease-associated Target Gene (TG): *ZNF589*



In this example, we utilize the B score values of TF-TF coordination for TG *ZNF589* in myelin (mSCs) Schwann Cells (SCs) where $-100 \leq B \leq 100$. The Tibial Nerve eQTL associated SNP rs9847953 has a very significant nominal p-value of 2e-119 with increases in *ZNF589* expression. Based on Pan UK Biobank Genome Wide Association Study (GWAS) summary statistics for the European population, we find that this SNP is associated with decreases (effect size = -0.1213, p = 2.104e-3) in injuries to the nervous system and decreases (effect size = -0.1885, p = 5.872e-4) in injury of nerves at wrist and hand level (Pan UK Biobank GWAS code: S64). Based on our GWAS-eQTL colocalization analysis, we find that posterior probability PP4 = 63.2% and posterior probability PP4 = 85.9% for *ZNF589* association with injuries to the nervous system and injury of nerves at wrist and hand level, respectively. Here, PP4 is the probability that the given SNP is associated and affects both traits: GWAS trait and TG expression trait (i.e. eQTL). NetREm predicts that this SNP is located within an intronic promoter for *ZNF589* and decreases bindings of repressor TFs like TEAD1 and ELF2 in mSCs and instead increases bindings of activator TFs KLF7, KLF15, ESRRA. There is a strong cooperation (i.e. positive coordination $B > 0$) of 14.49 between KLF7 and KLF15 in mSCs for regulating *ZNF589*. Across all final genes in mSCs, the average KLF15-TEAD1 coordination is 8.33 percentile and is a net negative interaction (i.e. overall antagonistic). This example shows how a common SNP may impact regulation of a TG in mSCs through regulatory mechanisms and TF-TF interaction networks. We also note that based on our predicted TF-TG regulatory network, the optimal coefficients c^* of the TFs for regulating TG *ZNF589* are approximately: TEAD1 (-0.000261), ESRRA (0.001142), KLF15 (0.000304), KLF7 (0.000395), ELF2 (-0.000064).

Figure B.20 Comparing Coordination Scores across Glial/Neuronal Cell-types and Conditions: Information Content Word Cloud representations of Cell-type TF-TF Coordination Network Links $-\mathbf{100} \leq \bar{\boldsymbol{B}} \leq \mathbf{100}$ and Comparison of Novel TF-TF Links with Physical Binding Support

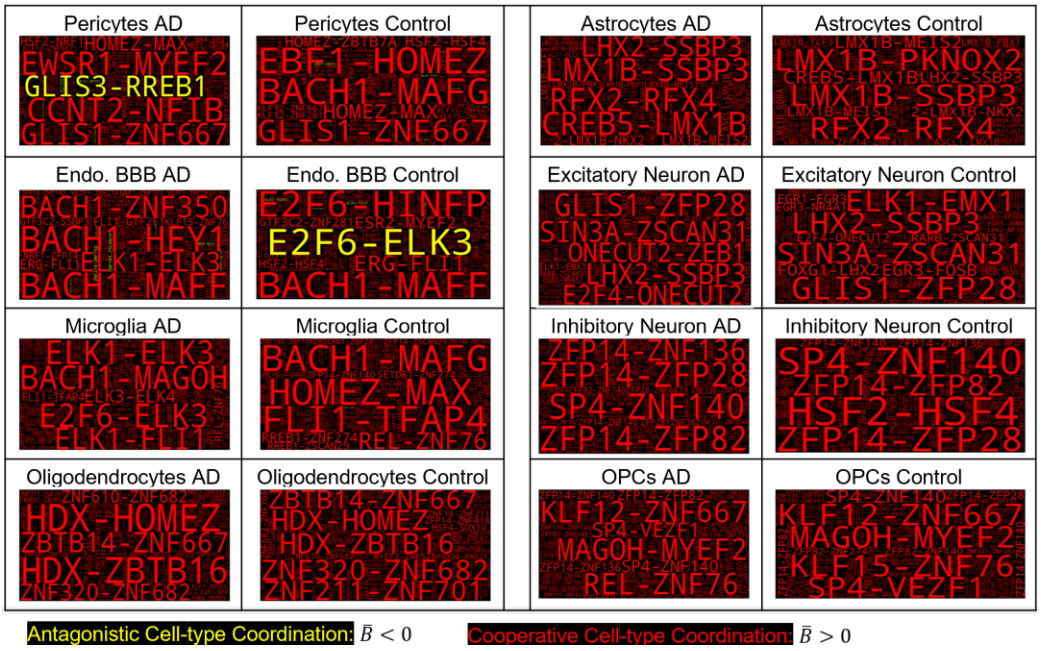


Figure B.20A) There are larger word sizes for larger $|\bar{B}|$. Here, red words have: $\bar{B} > 0$, yellow words have $\bar{B} < 0$. We show core links here (i.e. those with relatively high magnitudes of TF-TF coordination).

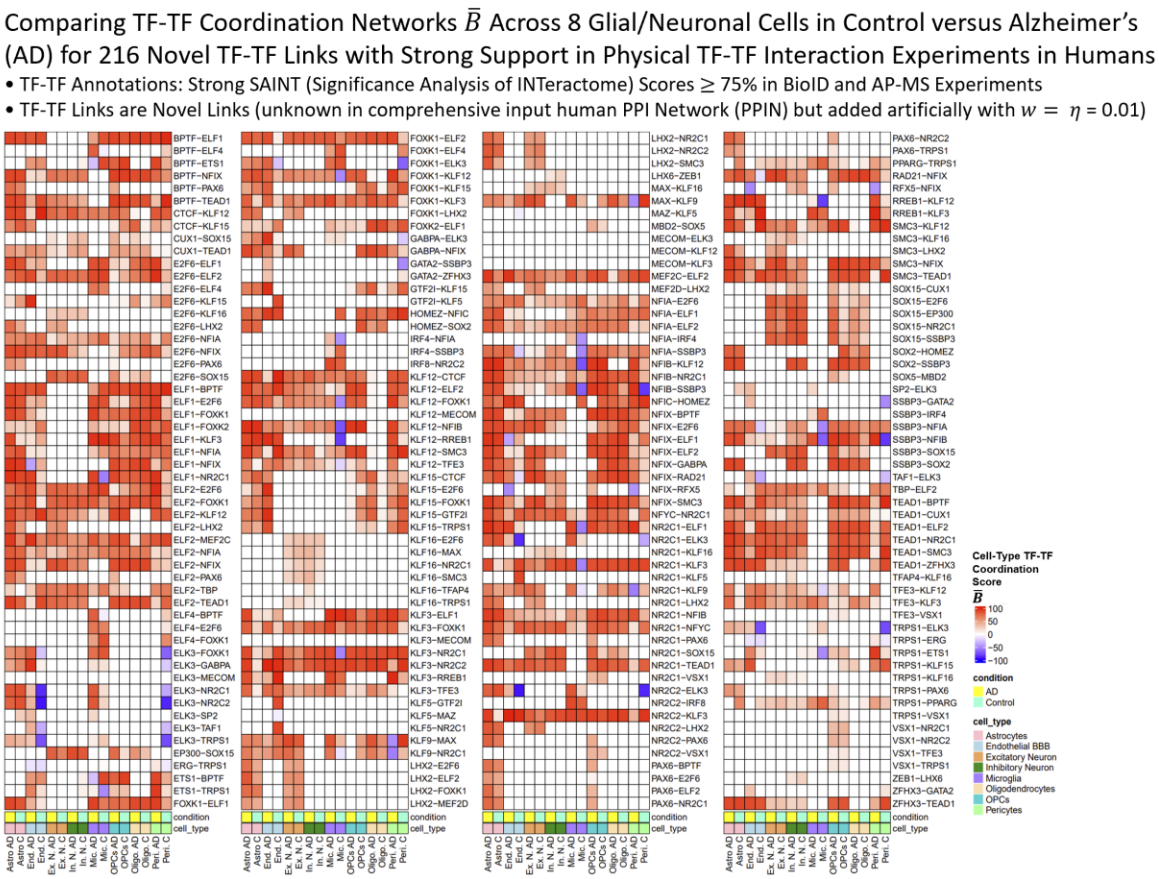


Figure B.20B) Comparative heatmap of 216 novel TF-TF coordination links (unknown in comprehensive input human PPIN but added artificially to make input PPIN fully-connected, with $w = \eta = 0.01$*)* that are found in at least 1 of the 16 different networks (8 cell-type networks, 2 conditions: Alzheimer's disease (AD) versus Controls). Here, these novel TF_i-TF_j links are discovered in (Göös et al. 2022), a recent study on direct, physical human TF interaction networks where these TFs are predicted to physically bind in AP-MS and/or BioID experiments with very strong SAINT scores of interactions (above 75%). Please note that other novel TF-TF links could correspond to indirect TF-TF interactions that are not yet known (since we used the most comprehensive version). Overall, this shows NetREm's novel TF-TF coordination links may be biologically meaningful discoveries and also shows differences between the cell-type TF-TF coordination scores $\bar{B}$ between human cell-types and/or conditions.

Figure B.21 Analyzing NetREm TF-TF predictions for Control Excitatory and Inhibitory Neurons based on Neural Cells for 6 TFs from UCSC Genome Browser: CTCF, EP300, EZH2, MXI1, RAD21, SMC3.

We use raw ENCODE3 ChIP-seq peaks TF clusters (**encRegTfbsClusteredWithCells.hg38.bed.gz**).

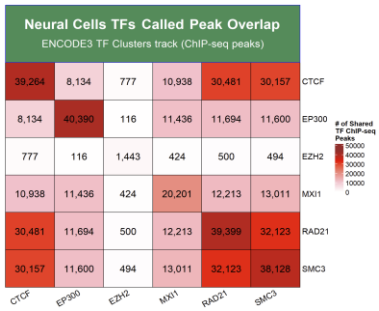


Figure B.21A) We see the peak overlap of TF Clusters shared among these 6 TFs.

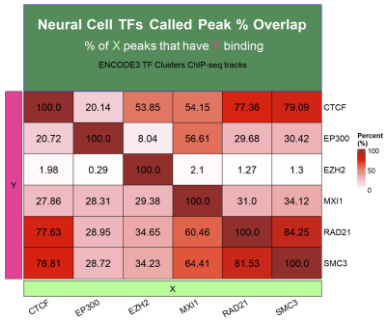


Figure B.21B) For instance, (X = 3: EZH2, Y = 2: EP300) refers to % of shared neural cell TF clusters peaks for EZH2 that have also have strong binding for EP300, which is ~8.04%. Conversely, (X = 2: EP300, Y = 3: EZH2) shows ~0.29% of peaks for EP300 have strong binding for EZH2.

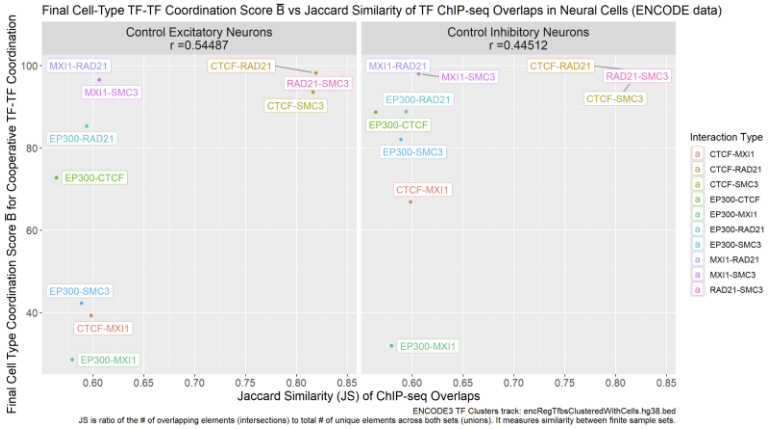


Figure B.21C) We calculate the Jaccard Similarity (JS) of these overlaps among the TF-TF pairs.

This JS(TF_i, TF_j) is based on: (# of common peaks for TF_i and TF_j) divided by the total # of unique peaks across TF_i and TF_j. So TF pairs with a strong JS will have a high overlap and this is a way to get 1 value for each TF-TF pair that quantifies similarity. This is on the x-axis. The y-axis has the cell-type TF-TF coordination score $\bar{B}$ and we note the correlations (r) across Excitatory Neurons (r = 0.544) and Inhibitory Neurons (r = 0.45) with Jaccard Similarity.

Figure B.22 Machine Learning Models to Predict Neurodegenerative Disease (class 1) or Not (class 0) TF-TF coordination links based on changes from Control to AD stages across 8 neuronal/glial cell-types.

We build classifier models (defaults from Python's sklearn package (Pedregosa et al. 2011)) for this task of predicting which TF-TF coordination links are annotated with neurodegenerative diseases. We use stratified 5-fold cross validation (CV). Each class is balanced: equal # of TF-TF links: 0 and 1.

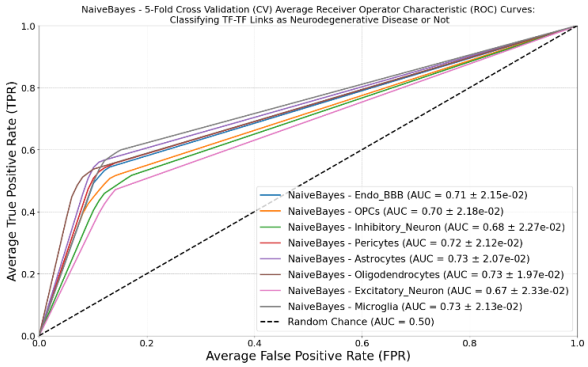


Figure B.22A) Naïve Bayes Classifier model

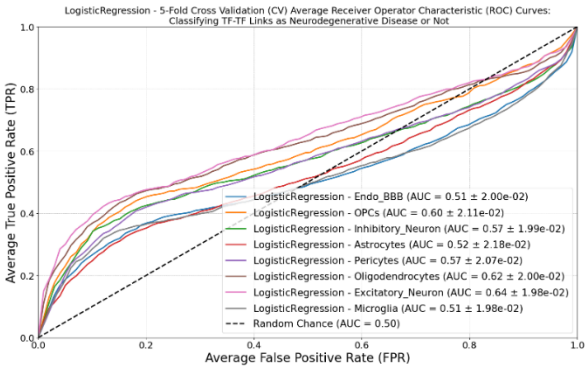


Figure B.22B) Logistic Regression classifier model

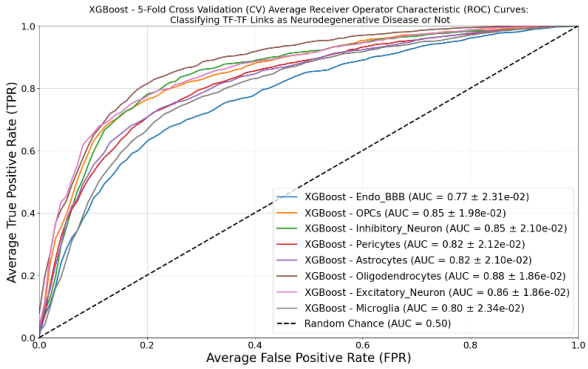


Figure B.22C) XGBoost classifier model

§ B.3 Supplementary tables

Table B.1: 1-Sided t-Test Coefficient Comparisons for Simulation Study

(NetREm versus Benchmarks)

The results of the 1-sided t-test comparison (assumption: equal variances = False) for raw coefficient values for the 5 Transcription Factor (TF) predictors (TF_1, TF_2, TF_3, TF_4, and TF_5) for the target gene (TG) y in the simulation study based on the following criteria: we run 1,000 simulations for $r(TF, TG)$ = [0.9, 0.5, 0.4, -0.3, -0.8] and 40% sparsity (0 values) for each variable in original training data, with Lasso model-type, $\alpha = 0.1$, and $\beta = 1$. In this case, we use all of the data for training (and the same exact data for testing) instead of holding anything out; thus, here our M = 10,000 cells. We standardize our data, so each variable has mean $\mu = 0$ and standard deviation of $\sigma = 1$. We retain the original simulated prior network from **Fig. 2A**. These results showcase NetREm's ability to statistically significantly increase the magnitudes of TF_2 (more positive) and TF_4 (more negative) and decrease the magnitude of TF_3 (making its value less positive), by incorporating the prior network as a constraint. This change in coefficient values for these 3 TFs does not appear to impact the coefficient values for major predictors TF_1 and TF_5. That is, in return, we retain our Null Hypothesis for TF_1 and TF_5 as the magnitude of their coefficients for TF_1 and TF_5 do not tend to decrease in any meaningful way.

1-Sided t Test Comparison: Coefficient Comparison for 1,000 Simulations										
TF Predictor	**Group 1**	**Group 2**	**Test statistic t**	**degrees of freedom (df)**	**p-value**	**Mean**		**Alternative Hypothesis H_A**	**Bounds**	
						Group 1	**Group 2**		**Low**	**Upper**
TF_1	NetREm	Lasso CV	-1,485.27	1,496.15	1	0.3772	0.6184	greater	-0.2414	Infinity
		Linear Regression	-1,484.88	1,495.73	1		0.6184		-0.2415	
		Ridge CV	-1,475.4	1,490.26	1		0.6181		-0.2411	
		ElasticNet CV	-1,483.75	1,498.90	1		0.6176		-0.2406	
TF_2		Lasso CV	958.011	1,561.34	0	0.2475	0.0883		0.1589	
		Linear Regression	954.5762	1,561.05	0		0.0888		0.1584	
		RidgeCV	954.3345	1,561.25	0		0.0889		0.1583	
		ElasticNet CV	958.2369	1,562.95	0		0.0885		0.1587	
TF_3		LassoCV	-204.604	1,468.69	0	0.0316	0.0632	less	- Infinity	-0.0314
		Linear Regression	-208.372	1,468.36	0		0.0638			-0.0320
		RidgeCV	-208.61	1,467.92	0		0.0639			-0.0320
		ElasticNet CV	-205.703	1,469.94	0		0.0634			-0.0315
TF_4		Lasso CV	-288.675	1,865.80	0	-0.0928	-0.0434			-0.0492
		Linear Regression	-284.77	1,865.46	0		-0.0440			-0.0485
		RidgeCV	-284.742	1,866.35	0		-0.0440			-0.0485

		ElasticNet CV	-288.374	1,867.02	0		-0.0435			-0.0491
TF_5		Lasso CV	543.5864	1,716.19	1	-0.1884	-0.2874			0.0992
		Linear Regression	545.2875	1,715.89	1		-0.2877			0.0996
		Ridge CV	546.054	1,716.69	1		-0.2878			0.0997
		ElasticNet CV	545.9327	1,719.20	1		-0.2876			0.0995

Table B.2: Standard Deviation of Coefficients in Simulation Study (NetREm versus Benchmarks)

Additional analysis of results for the NetREm simulation study for the same conditions mentioned in the description for **Table B.1.** The results of the robustness evaluation for NetREm versus benchmark models trained on data for 5 Transcription Factor (TF) predictors (TF_1, TF_2, TF_3, TF_4, TF_5) for the target gene (TG) with expression y in the simulation study. We use the following criteria: we run 1,000 simulations for $cor(TF, TG)$ = [0.9, 0.5, 0.4, -0.3, -0.8] and sparsity of 40% for the gene expression data (i.e. $\approx 40\%$ of the data for each of the 5 predictors in X and the response variable y are 0 values), with Lasso model-type and $\alpha = 0.1$ and $\beta = 1$. We simulate 10,000 cells and in each simulation, we simply vary the random seed that is used to generate the data (based on the underlying correlations). Please note that in this case we use all of the data for training (as well as for testing) without holding any data out, so our $M = 10{,}000$ cells. We retain the original simulated prior network from **Fig. 3.2A**. Results illustrate how NetREm standard deviations (variance of the models) tend to be lesser for each TF predictor relative to those from the other 4 benchmark models: ElasticNetCV, LassoCV, Linear Regression, and RidgeCV.

Standard Deviation of Coefficients in Demo Study after 1,000 simulations for $\beta = 1$					
Model	TF1	TF2	TF3	TF4	TF5
ElasticNetCV	0.004552	0.00459	0.004371	0.004308	0.004815
LassoCV	0.004567	0.004598	0.004378	0.004314	0.00483
LinearRegression	0.004569	0.004599	0.00438	0.004316	0.004831
NetREm	0.002353	0.002548	0.002184	0.003277	0.003137
RidgeCV	0.004598	0.004598	0.004382	0.004312	0.004827

Table B.3: Standard Deviation Across Different NetREm models

The results for the NetREm simulation study when running 1,000 simulations for $cor(TF, TG)$ = [0.9, 0.5, 0.4, -0.3, -0.8] where the data is ≈40% sparse, M = 7,000 (since 70% of the 10,000 cells in the gene expression data are used for training). We used the Lasso model-type and $\alpha = 0.1$ and β varied across 13 values: 0.1, 0.25, 0.5, 0.75, 1, 2.5, 5, 7.5, 10, 25, 50, 75, and 100. The data has been standardized for X and y data based on the training data. In each simulation, the pairwise correlations among the TF predictors vary. We retain the original simulated prior network from **Fig. 3.2A**. Results illustrate how NetREm standard deviations (variance of the models) tend to be lesser up to $\beta = 5$. Then they teeter a bit till $\beta = 50$. After which they plummet. After this, the increased strength of network penalization tends to worsen the performance of NetREm, which struggles to meet the network constraint while optimizing model performance (measured by Mean Square Error (MSE)). This illustrates the bias-variance trade-off.

Standard Deviation Values Across 1,000 Simulations for Fixed α and Varied β						
β	TF_1	TF_2	TF_3	TF_4	TF_5	Total Standard Deviation
0.1	0.00330	0.00366	0.00371	0.00329	0.00381	0.00329
0.25	0.00275	0.00313	0.00326	0.00346	0.00335	0.00275
0.5	0.00248	0.00279	0.00279	0.00333	0.00317	0.00248
0.75	0.00239	0.00263	0.00245	0.00329	0.00314	0.00239
1	0.00235	0.00255	0.00218	0.00328	0.00314	0.00218
2.5	0.00231	0.00241	0.00136	0.00326	0.00318	0.00136
5	0.00233	0.00239	0.00091	0.00325	0.00320	0.00091
7.5	0.00236	0.00240	0.00076	0.00324	0.00320	0.00076
10	0.00241	0.00245	0.00070	0.00328	0.00325	0.00070
25	0.00254	0.00257	0.00062	0.00327	0.00326	0.00062
50	0.00090	0.00093	0.00017	0.00061	0.00063	0.00017
75	0.00059	0.00062	0	0	0	0
100	0.00051	0.00054	0	0	0	0

Table B.4: Metrics for Predicting TF-TG regulatory links for human Hematopoietic Stem Cells (HSCs)

We vary $\beta \in \{0.5, 1, 5, 10\}$ and use LassoCV. As β increases, overall accuracy dips initially but improves to 79% at $\beta = 10$, with sensitivity peaking early before declining; specificity exhibits an inverse pattern.

Metrics for Predicting TF-TG Regulatory Links for Human Hematopoietic Stem Cells (HSCs)													
Goal:	?	↑	↑	↑	↓	↓	↓	↑	↑	↑	↑	↑	↑
Model	# of Rows (TF-TG)	True Positives	True Negatives	Total Correct (TP + TN)	False Negatives	False Positives	Total Errors (FN + FP)	Overall Accuracy	Sensitivity (recall)	Specificity	Balanced Accuracy (BACC)	Precision	F1 Score
NetREm (β = 1, α = 0.01)	1,133,617	181,624	663,909	845,533	127,011	951,993	1,079,004	43.9%	58.85%	41.09%	49.97%	16.02%	25.2%
NetREm (β = 0.5, LassoCV)	33,046	28,541	1,465,836	1,494,377	280,094	150,066	430,160	77.6%	9.25%	90.71%	49.98%	15.98%	11.7%
NetREm (β = 1, LassoCV)	187,230	30,379	1,459,051	1,489,430	278,256	156,851	435,107	77.4%	9.84%	90.29%	50.07%	16.23%	12.3%
NetREm (β = 5, LassoCV)	137,220	48,796	1,358,689	1,407,485	259,839	257,213	517,052	73.1%	15.81%	84.08%	49.95%	15.95%	15.9%
NetREm (β = 10, LassoCV)	178,607	21,288	1,499,970	1,521,258	287,347	115,932	403,279	79.0%	6.90%	92.83%	49.86%	15.51%	9.5%
GRNBoost2	741,189	133,696	1,008,409	1,142,105	174,939	607,493	782,432	59.3%	43.32%	62.41%	52.86%	18.04%	25.5%
ElasticNetCV	3,837,442	6,302	1,586,782	1,593,084	302,333	29,120	331,453	82.8%	2.04%	98.20%	50.12%	17.79%	3.7%
LassoCV	1,884,487	5,877	1,588,733	1,594,610	302,758	27,169	329,927	82.9%	1.90%	98.32%	50.11%	17.78%	3.4%
Linear Regression	35,422	303,056	34,471	337,527	5,579	1,581,431	1,587,010	17.5%	98.19%	2.13%	50.16%	16.08%	27.6%
RidgeCV	1,884,487	303,056	34,471	337,527	5,579	1,581,431	1,587,010	17.5%	98.19%	2.13%	50.16%	16.08%	27.6%

Table B.5: Metrics for Predicting TF-TG Regulatory Links for Mouse Embryonic Stem Cells (mESCs)

Metrics for Predicting TF-TG Regulatory Links for Mouse Embryonic Stem Cells (mESCs)													
Goal:	?	↑	↑	↑	↓	↓	↓	↑	↑	↑	↑	↑	↑
Model	# of Rows (TF-TG)	True Positives	True Negatives	Total Correct (TP + TN)	False Negatives	False Positives	Total Errors (FN + FP)	Overall Accuracy	Sensitivity (recall)	Specificity	Balanced Accuracy (BACC)	Precision	F1 Score
NetREm (β = 1, α = 0.05, old STRINGv11 PPI)	466,220	45,451	3,030,492	3,075,943	251,969	420,769	672,738	82.05%	15.28%	87.81%	51.5%	9.75%	11.9%
NetREm (β = 1, α = 0.05, new STRINGv12 PPI)	466,654	45,423	3,030,030	3,075,453	251,997	421,231	673,228	82.04%	15.27%	87.79%	51.5%	9.73%	11.9%
GRNBoost2	1,634,683	156,349	1,972,927	2,129,276	141,071	1,478,334	1,619,405	56.80%	52.57%	57.17%	54.9%	9.56%	16.2%
ElasticNetCV	232,211	23,021	3,242,071	3,265,092	274,399	209,190	483,589	87.10%	7.74%	93.94%	50.8%	9.91%	8.7%
LassoCV	213,819	21,275	3,258,717	3,279,992	276,145	192,544	468,689	87.50%	7.15%	94.42%	50.8%	9.95%	8.3%
Linear Regression	3,029,325	268,632	690,568	959,200	28,788	2,760,693	2,789,481	25.59%	90.32%	20.01%	55.2%	8.87%	16.1%
RidgeCV	2,982,912	266,479	734,828	1,001,307	30,941	2,716,433	2,747,374	26.71%	89.60%	21.29%	55.4%	8.93%	16.2%

Results where there are 3,748,681 total possible TF-TG regulatory links. NetREm's sensitivity is higher than that for ElasticNet and Lasso.

Table B.6: Metrics for Predicting TF-TG Regulatory Links for Normalized Mouse Dendritic Cells (mDCs)

We fix $\beta = 1$ and alter α for 7 values, noting improved accuracy, precision, and specificity and drops in F1 score, sensitivity, balanced accuracy (BACC). TP = True Positives ; TN = True Negatives ; FN = False Negatives; FP = False Positives.

Metrics for Predicting TF-TG Regulatory Links for Normalized Mouse Dendritic Cells (mDCs)													
Goal:	?	↑	↑	↑	↓	↓	↓	↑	↑	↑	↑	↑	↑
Model	# of Rows (TF-TG)	TPs	TNs	Total Correct (TP + TN)	FNs	FPs	Total Errors (FN + FP)	Overall Accuracy	Sensitivity (recall)	Specificity	Balanced Accuracy (BACC)	Precision	F1 Score
NetREm (β = 1, α = 0.01)	640,144	135,941	164,070	300,011	40,784	504,203	544,987	35.50%	76.92%	24.55%	50.74%	21.24%	33.28%
NetREm (β = 1, α = 0.025)	398,159	88,617	358,731	447,348	88,108	309,542	397,650	52.94%	50.14%	53.68%	51.91%	22.26%	30.83%
NetREm (β = 1, α = 0.05)	325,950	73,704	416,027	489,731	103,021	252,246	355,267	57.96%	41.71%	62.25%	51.98%	22.61%	29.32%
NetREm (β = 1, α = 0.075)	79,671	22,753	611,355	634,108	153,972	56,918	210,890	75.04%	12.87%	91.48%	52.18%	28.56%	17.75%

NetREm ($\beta = 1,\ \alpha = 0.1$)	38,148	12,162	642,287	654,449	164,563	25,986	190,549	77.45%	6.88%	96.11%	51.50%	31.88%	11.32%
NetREm ($\beta = 1,\ \alpha = 0.125$)	20,412	6,943	654,804	661,747	169,782	13,469	183,251	78.31%	3.93%	97.98%	50.96%	34.01%	7.04%
NetREm ($\beta = 1,\ \alpha = 0.15$)	12,235	4,145	660,183	664,328	172,580	8,090	180,670	78.62%	2.35%	98.79%	50.57%	33.88%	4.39%
NetREm ($\beta = 1$, LassoCV)	109,647	24,887	583,513	608,400	151,838	84,760	236,598	72.00%	14.08%	87.32%	50.70%	22.70%	17.38%
NetREm $\beta = 0.1$, LassoCV)	142,050	32,458	558,681	591,139	144,267	109,592	253,859	69.96%	18.37%	83.60%	50.98%	22.85%	20.36%
NetREm ($\beta = 10$, LassoCV)	60,019	13,817	622,071	635,888	162,908	46,202	209,110	75.25%	7.82%	93.09%	50.45%	23.02%	11.67%
Scribe	427,042	170,986	412,217	583,203	5,739	256,056	261,795	69.02%	96.75%	61.68%	79.22%	40.04%	56.64%
Pearson mean	185,328	88,185	571,130	659,315	88,540	97,143	185,683	78.03%	49.90%	85.46%	67.68%	47.58%	48.71%
PIDC mean	314,028	89,366	443,611	532,977	87,359	224,662	312,021	63.07%	50.57%	66.38%	58.47%	28.46%	36.42%
MERLIN	36,492	18,414	650,195	668,609	158,311	18,078	176,389	79.13%	10.42%	97.29%	53.86%	50.46%	17.27%
SCENIC	32,771	17,962	653,464	671,426	158,763	14,809	173,572	79.46%	10.16%	97.78%	53.97%	54.81%	17.15%
Inferelator	23,649	10,096	654,720	664,816	166,629	13,553	180,182	78.68%	5.71%	97.97%	51.84%	42.69%	10.08%
knnDREMI	145,438	70,594	593,429	664,023	106,131	74,844	180,975	78.58%	39.95%	88.80%	64.37%	48.54%	43.83%
LEAP mean	258,009	66,184	476,448	542,632	110,541	191,825	302,366	64.22%	37.45%	71.30%	54.37%	25.65%	30.45%
SCODE mean	156,829	64,447	575,891	640,338	112,278	92,382	204,660	75.78%	36.47%	86.18%	61.32%	41.09%	38.64%
SILGGM mean	156,829	64,447	575,891	640,338	112,278	92,382	204,660	75.78%	36.47%	86.18%	61.32%	41.09%	38.64%
ElasticNetCV	94,552	23,564	597,285	620,849	153,161	70,988	224,149	73.47%	13.33%	89.38%	51.36%	24.92%	17.37%
Linear Regression	521,916	109,134	255,491	364,625	67,591	412,782	480,373	43.15%	61.75%	38.23%	49.99%	20.91%	31.24%
RidgeCV	522,009	109,159	255,423	364,582	67,566	412,850	480,416	43.15%	61.77%	38.22%	49.99%	20.91%	31.24%
LassoCV	89,292	22,288	601,269	623,557	154,437	67,004	221,441	73.79%	12.61%	89.97%	51.29%	24.96%	16.76%

Table B.7 Reference table for Magnitude of TF-TF Coordination Score Group $|\bar{B}|$ in mouse Embryonic Stem Cells (mESCs)

The mouse STRINGdb version 11 (v11) refers to the protein-protein interaction (PPI) network for mice as of January 19, 2019 to October 17, 2020. The later release of this PPI (and current version as of the timepoint this manuscript was written) is version 12 (v12), which includes data since July 26, 2023.

Reference Table: Magnitude of TF-TF Coordination Score $\|\bar{B}\|$ Group Comparisons in Mouse Embryonic Stem Cells (mESCs)					
Goal	**Group 1**	**Group 2**	**Goal #**	# of TF-TF Coordination Links in Group 1	# of TF-TF Coordination Links in Group 2

known TF-TF links (in v11) versus links that are artificial (in v11)	PPI links in v11	PPI links not in v11	1	10,088	27,742
known links (in v11): comparing TF-TF links that are retained to those that are removed (in v12)	PPI links in v11 and v12	PPI links in v11 that are not in v12	2	8,590	1,498
artificial links (not in v11): comparing valid discovery links (found in v12) with those still not found in v12	PPI links not in v11 but found in v12	PPI links found in 0 networks (not in v11 and not in 12)	3	5,122	22,620
comparing valid discovery links (not in v11 but in v12) with removed links (False Positives; in v11 but not in v12)	PPI links not in v11 but found in v12	PPI links in v11 that are not in v12	4	5,122	1,498
comparing known links (any links in v11) with valid discovery links (not in v11 but in v12)	all PPI links found in v11	PPI links not in v11 but found in v12	5	10,088	5,122
comparing known and retained links (in v11 and v12) with valid discovery links (not in v11 but in v12)	PPI links in v11 and v12	PPI links not in v11 but found in v12	6	8,590	5,122
comparing unknown links (not in any network) with removed links (from v11 to v12)	PPI links found in 0 networks (not in v11 and not in 12)	PPI links in v11 that are not in v12	7 and 8	22,620	1,498

For other applications, please focus on just the first 4 columns: Goal, Group 1, Group 2, Goal #.

Table B.8: 1-sided t-test comparison of magnitude of TF-TF coordination scores $|\bar{B}|$ in Mouse Embryonic Stem Cells (mESCs)

1-Sided t Test Comparison: Magnitude of TF-TF Coordination Scores $\|\bar{B}\|$ in Mouse Embryonic Stem Cells (mESCs)												
						significance (p)		Mean			95% Confidence Bounds	
Goal #	Metric	Test statistic t	degrees of freedom (df)	p-value	p-adj	p	p-adj	Group 1	Group 2	Alt. Hypothesis	Low	Upper
1	Percentile (Average Coordination Score)	22.106	17,806.89	3.79E-107	6.07E-106	Yes	Yes	54.061	46.46	greater	7.04	Inf
2		11.311	2,103.66	3.85E-29	5.00E-28	Yes	Yes	55.408	46.34	greater	7.75	Inf
3		10.905	7,409.43	8.85E-28	9.74E-27	Yes	Yes	50.606	45.52	greater	4.322	Inf
4		5.0258	2,577.99	2.68E-07	1.61E-06	Yes	Yes	50.606	46.34	greater	2.87	Inf
5		6.6829	10,070.31	1.23E-11	9.87E-11	Yes	Yes	54.061	50.61	greater	2.605	Inf
6		9.0348	10,553.28	9.65E-20	9.65E-19	Yes	Yes	55.408	50.61	greater	3.928	Inf

7		-1.082	1,712.14	8.60E-01	1.00E+00	No	No	45.516	46.34	greater	-2.075	Inf
8		-1.082	1,712.14	1.40E-01	5.58E-01	No	No	45.516	46.34	less	-Inf	0.428
1	|Average Coordination Score B|	17.878	14,312.13	5.14E-71	7.72E-70	Yes	Yes	0.0104	0.007	greater	0.003	Inf
2		12.236	2,835.46	6.88E-34	9.63E-33	Yes	Yes	0.0111	0.006	greater	0.004	Inf
3		10.937	6,604.75	6.63E-28	7.96E-27	Yes	Yes	0.009	0.006	greater	0.002	Inf
4		6.5928	3,107.11	2.53E-11	1.77E-10	Yes	Yes	0.009	0.006	greater	0.002	Inf
5		4.7551	11,355.06	1.00E-06	5.02E-06	Yes	Yes	0.0104	0.009	greater	9E-04	Inf
6		6.8049	12,034.56	5.30E-12	4.77E-11	Yes	Yes	0.0111	0.009	greater	0.002	Inf
7		-0.118	1,706.17	5.47E-01	1.00E+00	No	No	0.0062	0.006	greater	-6E-04	Inf
8		-0.118	1,706.17	4.53E-01	1.00E+00	No	No	0.0062	0.006	less	-Inf	5E-04

Alt. Hypothesis is Alternative Hypothesis H_A and refers to Group 1 versus Group 2. Inf is Infinity.

- Novel discoveries (not in V11; in V12) than TF-TF links that are removed (in V11 but not in V12 networks). → NetREm prioritizes True Positives (TPs) to False Positives (FPs) and may be predictive of truly meaningful future TF-TF links.
- Novel discoveries than in TF-TF links that remain unknown (in neither network)
- In V11 networks than links not in V11 network → NetREm network prioritizes known links
- In both networks than novel discoveries → NetREm still prioritizes known links.
- In V11 (may be removed/retained in V12) than novel discoveries → NetREm prioritizes known links.
- In both networks than removed links → NetREm prioritizes potential TPs to FPs

Table B.9 Significance Tests from 1-sided t-tests of TF-TF Coordination Scores in Mouse Dendritic Cells (mDCs)

1-Sided t-Test Significance Results (P-adj < 0.05) of magnitude of TF-TF coordination scores $\|\overline{B}\|$ for Mouse Dendritic Cells Based on Input Mouse STRINGdb V11 Protein-Protein Interaction (PPI) Network (Older)								
	Goal # (1 = significant, 0 = not significant)							
Varying α for the NetREm model (for β = 1)	**1**	**2**	**3**	**4**	**5**	**6**	**8**	Total # of Tests (Goals) the Model is Significant For
0.01	1	0	1	0	1	1	1	5
0.025	1	1	1	0	1	1	1	6
0.05	1	1	1	0	1	1	1	6
0.075	1	1	1	0	1	1	1	6
0.1	1	1	1	1	1	1	1	7
0.125	1	1	1	1	0	1	0	5
0.15	1	1	1	1	0	1	0	5
CV	1	1	1	0	1	1	1	6
# of models that are significant for goal	**8**	**7**	**8**	**3**	**6**	**8**	**6**	**46**

P-adj is taken internally for each model. Altering the α for mDCs leads to additional statistically significant findings. Across these different NetREm models, we still find $|\bar{B}|$ is significantly higher for TF-TF links that are known (in V11) versus those that are not. Similarly, we find $|\bar{B}|$ is significantly higher for TF-TF links that are known and retained in both networks than for valid discoveries, which again shows how NetREm prioritizes known links. Nonetheless, we do find that NetREm's $|\bar{B}|$ is significantly higher for TF-TF links that are eventually discovered in V12 (valid discoveries) than those that still remain unknown in V12. This is a strong sign, showing NetREm's ability to prioritize future valid discoveries (with higher magnitude of coordination scores) to links that may be false positive (or at least not discovered yet). As we alter α, we do find significant results for additional 1-sided tests among the groups. This illustrates how NetREm may utilize an older PPI network and single-cell gene expression relationships to potentially uncover and predict novel relationships. There is a form of transfer learning.

Table B.10: 1-sided t-test comparison of Magnitude of TF-TF Coordination Scores $|\bar{B}|$ in Mouse Dendritic Cells (mDCs)

1-Sided t Test Comparison: Magnitude of TF-TF Coordination Scores in Mouse Dendritic Cells (mDCs)												
Goal #	α	Test statistic t	degrees of freedom (df)	p-value	p-adj	significance (p)		Mean		Alt. Hypothesis	95% Confidence Bounds	
						p	p-adj	Group 1	Group 2		Low	Upper
1	LassoCV	19.08	2,741.03	1.35E-76	1.08E-75	Yes	Yes	61.36	47.13	greater	13.002	Inf
2		2.85	309.69	0.00233	0.006989	Yes	Yes	62.06	56.73	greater	2.244	Inf
3		3.55	1,046.25	0.000199	0.000798	Yes	Yes	50.61	46.64	greater	2.1274	Inf
4		-3.03	412.36	0.998687	1	No	No	50.61	56.73	greater	-9.464	Inf
5		8.63	1,510.57	7.36E-18	4.42E-17	Yes	Yes	61.36	50.61	greater	8.7014	Inf
6		8.99	1,591.57	3.32E-19	2.32E-18	Yes	Yes	62.06	50.61	greater	9.3583	Inf
7		-5.72	247.6	1	1	No	No	46.64	56.73	greater	-13.01	Inf
8		-5.72	247.6	1.56E-08	7.80E-08	Yes	Yes	46.64	56.73	less	-Inf	-7.18
1	0.01	19.64	2,715.33	1.03E-80	8.24E-80	Yes	Yes	61.76	47.03	greater	13.498	Inf
2		1.81	318.36	0.035805	0.107415	Yes	No	62.2	58.9	greater	0.288	Inf
3		2.37	1,046.99	0.008903	0.035612	Yes	Yes	49.34	46.7	greater	0.8078	Inf
4		-4.84	423.23	0.999999	1	No	No	49.34	58.9	greater	-12.82	Inf
5		9.98	1,527.44	4.38E-23	2.63E-22	Yes	Yes	61.76	49.34	greater	10.372	Inf
6		10.08	1,618.12	1.67E-23	1.17E-22	Yes	Yes	62.2	49.34	greater	10.755	Inf
7		-7.12	248.86	1	1	No	No	46.7	58.9	greater	-15.02	Inf
8		-7.12	248.86	5.66E-12	2.83E-11	Yes	Yes	46.7	58.9	less	-Inf	-9.37
1	0.025	18.37	2,729.93	1.86E-71	1.49E-70	Yes	Yes	60.99	47.22	greater	12.529	Inf
2		2.65	314.65	0.004276	0.012827	Yes	Yes	61.63	56.75	greater	1.8359	Inf
3		3.67	1,055.59	0.00013	0.000518	Yes	Yes	50.76	46.73	greater	2.2192	Inf
4		-3.02	412.62	0.998642	1	No	No	50.76	56.75	greater	-9.272	Inf
5		8.31	1,540.92	1.08E-16	6.46E-16	Yes	Yes	60.99	50.76	greater	8.2004	Inf
6		8.61	1,627.77	8.20E-18	5.74E-17	Yes	Yes	61.63	50.76	greater	8.7941	Inf

7		-5.78	248.53	1	1	No	No	46.73	56.75	greater	-12.89	Inf
8		-5.78	248.53	1.13E-08	5.64E-08	Yes	Yes	46.73	56.75	less	-Inf	-7.16
1	0.05	17.6	2,753.82	4.16E-66	3.32E-65	Yes	Yes	60.47	47.35	greater	11.895	Inf
2		3.08	315.83	0.00114	0.003421	Yes	Yes	61.21	55.62	greater	2.5954	Inf
3		5.12	1,055.27	1.84E-07	9.04E-07	Yes	Yes	52.3	46.66	greater	3.8269	Inf
4		-1.68	419.31	0.953342	1	No	No	52.3	55.62	greater	-6.562	Inf
5		6.63	1,526.78	2.26E-11	1.36E-10	Yes	Yes	60.47	52.3	greater	6.1442	Inf
6		7.06	1,613.31	1.23E-12	8.58E-12	Yes	Yes	61.21	52.3	greater	6.8339	Inf
7		-5.23	249.26	1	1	No	No	46.66	55.62	greater	-11.78	Inf
8		-5.23	249.26	1.81E-07	9.04E-07	Yes	Yes	46.66	55.62	less	-Inf	-6.13
1	0.075	16	2,706.86	2.06E-55	1.65E-54	Yes	Yes	59.66	47.54	greater	10.871	Inf
2		2.83	310.55	0.002498	0.007495	Yes	Yes	60.37	55	greater	2.2351	Inf
3		5.6	1,044.45	1.40E-08	8.40E-08	Yes	Yes	53.05	46.77	greater	4.427	Inf
4		-0.96	408.54	0.830053	1	No	No	53.05	55	greater	-5.334	Inf
5		5.27	1,529.02	7.95E-08	3.97E-07	Yes	Yes	59.66	53.05	greater	4.5474	Inf
6		5.7	1,611.50	7.26E-09	5.08E-08	Yes	Yes	60.37	53.05	greater	5.2076	Inf
7		-4.6	247.02	0.999997	1	No	No	46.77	55	greater	-11.19	Inf
8		-4.6	247.02	3.43E-06	1.37E-05	Yes	Yes	46.77	55	less	-Inf	-5.27
1	0.1	16.91	2,689.28	2.34E-61	1.87E-60	Yes	Yes	55.17	41.12	greater	12.68	Inf
2		5.67	299.46	1.66E-08	9.99E-08	Yes	Yes	56.8	44.46	greater	8.7477	Inf
3		9.42	1,064.39	1.36E-20	9.53E-20	Yes	Yes	50.84	39.76	greater	9.1412	Inf
4		2.76	373.41	0.003023	0.009068	Yes	Yes	50.84	44.46	greater	2.5692	Inf
5		3.25	1,597.97	0.000587	0.00235	Yes	Yes	55.17	50.84	greater	2.1369	Inf
6		4.39	1,659.70	6.10E-06	3.05E-05	Yes	Yes	56.8	50.84	greater	3.723	Inf
7		-2.27	244.77	0.988081	0.988081	No	No	39.76	44.46	greater	-8.11	Inf
8		-2.27	244.77	0.011919	0.023838	Yes	Yes	39.76	44.46	less	-Inf	-1.29
1	0.125	18.65	2,425.85	6.32E-73	5.05E-72	Yes	Yes	39.78	22.43	greater	15.823	Inf
2		8.02	325.16	9.46E-15	5.68E-14	Yes	Yes	42.16	24.16	greater	14.295	Inf
3		13.61	1,005.01	4.06E-39	2.84E-38	Yes	Yes	37.75	20.28	greater	15.358	Inf
4		5.7	405.04	1.18E-08	5.92E-08	Yes	Yes	37.75	24.16	greater	9.6571	Inf
5		1.36	1,654.28	0.087545	0.17509	No	No	39.78	37.75	greater	-0.432	Inf
6		2.88	1,731.16	0.002038	0.008152	Yes	Yes	42.16	37.75	greater	1.8837	Inf
7		-1.86	242.98	0.968227	0.968227	No	No	20.28	24.16	greater	-7.319	Inf
8		-1.86	242.98	0.031773	0.09532	Yes	No	20.28	24.16	less	-Inf	-0.44
1	0.15	17	2,217.54	2.59E-61	2.07E-60	Yes	Yes	26.68	11.82	greater	13.418	Inf
2		9.29	381.76	5.81E-19	3.48E-18	Yes	Yes	28.95	11.72	greater	14.173	Inf
3		11.74	956.82	3.99E-30	2.79E-29	Yes	Yes	23.93	10.12	greater	11.873	Inf
4		6.16	474	7.58E-10	3.79E-09	Yes	Yes	23.93	11.72	greater	8.9483	Inf
5		1.95	1,708.62	0.025401	0.076202	Yes	No	26.68	23.93	greater	0.433	Inf
6		3.46	1,810.92	0.000274	0.001096	Yes	Yes	28.95	23.93	greater	2.6322	Inf
7		-0.97	242.78	0.832868	0.832868	No	No	10.12	11.72	greater	-4.32	Inf

8		-0.97	242.78	0.167132	0.334263	No	No	10.12	11.72	less	-Inf	1.13

Table B.11 Significant 1-Sided Welch t-Test Comparisons for TF-TF Coordination Score Groups in Human PBMCs

1-Sided Welch t Test Comparison: Magnitude of TF-TF Coordination Scores in Human Peripheral Blood Mononuclear Cells (PBMCs) for Results that are Significant (P-adj < 0.05)											
Goal #	Cell-type	Test statistic t	degrees of freedom (df)	p-adj	# of Values		Mean		Alt. Hypothesis	95% Confidence Bounds	
					Group 1	Group 2	Group 1	Group 2		Low	Upper
1	Naive CD4 T cell (C0)	83.9	61,975.6	0	54,880	875,380	57.49	46.31	greater	10.96	Inf
5		56.8	117,656.1	0	54,880	65,914	57.49	47.46	greater	9.75	Inf
6		47.5	69,934.5	0	34,766	65,914	57.26	47.46	greater	9.46	Inf
8		-57.5	21,262.0	0	809,466	20,114	46.21	57.90	less	-Inf	-11.35
1	CD14 Mono cell (C1)	145.9	60,949.1	0	50,930	749,200	65.35	48.72	greater	16.45	Inf
5		86.4	109,860.2	0	50,930	59,210	65.35	51.11	greater	13.97	Inf
6		83.1	78,500.9	0	33,044	59,210	66.58	51.11	greater	15.16	Inf
8		-82.8	19,362.9	0	689,990	17,886	48.51	63.09	less	-Inf	-14.29
1	Memory CD4 T cell (C2)	103.4	59,280.3	0	51,094	796,226	61.68	49.19	greater	12.29	Inf
2		2.7	41,491.6	9.8E-03	32,414	18,680	61.92	61.27	greater	0.25	Inf
5		68.5	112,358.1	0	51,094	62,244	61.68	50.37	greater	11.04	Inf
6		60.6	70,252.4	0	32,414	62,244	61.92	50.37	greater	11.23	Inf
8		-65.5	19,967.3	0	733,982	18,680	49.09	61.27	less	-Inf	-11.88
1	B cell (C3)	137.8	53,771.0	0	45,802	750,754	65.54	49.03	greater	16.31	Inf
5		92.2	104,436.5	0	45,802	59,366	65.54	50.17	greater	15.09	Inf
6		85.8	67,922.6	0	29,678	59,366	66.43	50.17	greater	15.95	Inf
8		-79.8	17,282.1	0	691,388	16,124	48.94	63.89	less	-Inf	-14.65
1	CD8 T cell (C4)	126.1	53,680.6	0	45,680	720,820	64.45	49.08	greater	15.17	Inf
5		86.8	102,360.0	0	45,680	57,028	64.45	49.68	greater	14.49	Inf
6		80.2	66,818.3	0	29,312	57,028	65.23	49.68	greater	15.23	Inf
8		-73.9	17,567.7	0	663,792	16,368	49.03	63.04	less	-Inf	-13.70
1	FCGR3A Mono cell (C5)	72.0	50,422.4	0	44,170	599,836	59.81	49.28	greater	10.29	Inf
5		48.3	92,535.7	0	44,170	49,816	59.81	50.46	greater	9.03	Inf
8		-56.6	16,107.7	0	550,020	15,146	49.17	61.69	less	-Inf	-12.15
1	Natural Killer cell (C6)	101.0	48,949.9	0	42,448	595,154	63.34	49.05	greater	14.06	Inf
5		69.7	89,625.6	0	42,448	47,786	63.34	50.02	greater	13.01	Inf
6		56.4	55,629.4	0	27,160	47,786	62.67	50.02	greater	12.28	Inf
8		-75.4	16,417.7	0	547,368	15,288	48.96	64.55	less	-Inf	-15.24
4	Platelet cell (C8)	3.4	2,148.2	2.9E-03	2,938	1,134	50.87	47.68	greater	1.64	Inf

7		3.1	1,297.8	6.5E-03	19,670	1,134	50.23	47.68	greater	1.21	Inf

Table B.12: Significance Tests from 1-sided t-tests of TF-TF Coordination Scores in human Peripheral Blood Mononuclear Cells (PBMCs)

1-Sided t-Test Significance Results (P-adj < 0.05) of magnitude of TF-TF coordination scores $\|\bar{B}\|$ for Human Peripheral Blood Mononuclear Cells Based on Input Human STRINGdb V11 Protein-Protein Interaction (PPI) Network (Older)								
NetREm run with b = 1, a from LassoCV	**Goal #** (1 = significant, 0 = not significant)							Total # of Tests (Goals) the Cell-Type is Significant For
Cell-type	**1**	**2**	**4**	**5**	**6**	**7**	**8**	
B_C3	1			1	1		1	4
CD14_Mono_C1	1			1	1		1	4
CD8_T_C4	1			1	1		1	4
FCGR3A_Mono_C5	1			1			1	3
Memory_CD4_T_C2	1	1		1	1		1	5
Naive_CD4_T_C0	1			1	1		1	4
NaturalKiller_C6	1			1	1		1	4
Platelet_C8			1			1		2
# of cell-types that are significant for goal	**7**	**1**	**1**	**7**	**6**	**1**	**7**	**30**

Table B.13: Relative Percentiles of TF activity for core Schwann cell TFs in terms of # of TGs they regulate

Metrics for the 8 core Schwann Cell (SC) TFs that we have additional experimental data for. Percentile (%) is between 0 and 1 and represents how many other SC-subtype TFs that given core TF is ahead of in terms of the # of TGs it is predicted to regulate. For instance, EGR2 is the 90.5% percentile, roughly, as NetREm's final model predicts it regulates 1,522 TGs in myelinating (mSCs) human Schwann cells (SCs); on average, the 221 TFs in mSCs tend to regulate around 829.15 TGs.

TF	mSC TGs	nmSC TGs	percentile (mSC)	percentile (nmSC)
EGR2	1,522	0	0.904977	0
NR2F2	508	1,239	0.285068	0.484649
RXRG	193	1,556	0.144796	0.649123
SOX10	1,003	1,472	0.613122	0.587719
SREBF1	1,004	1,610	0.61991	0.679825
STAT1	1,337	1,895	0.832579	0.842105
TEAD1	1,878	2,527	0.995475	0.969298
YY1	1,329	2,015	0.81448	0.907895

Global Metrics across all TFs in the Human Schwann Cell (SCs)

	mSC TGs	nmSC TGs
# of TFs	221	228
Average	829.15	1,217.28
Median	839	1,279
Min	1	1
Max	2,126	3,188

We note that these core SC TFs tend to have relatively meaningful percentiles (in terms of the # of TGs they are predicted to regulate). In particular, TEAD1 is at the 99th and 97th percentile in mSCs and nmSCs, respectively. The lowest percentile is for RXRG in mSCs (14.5%), but we do show RXRG has stronger behavior in nmSCs (of 65th percentile). This may be attributed to RXRG's preferential expression in nmSCs over mSCs based on findings from single-cell data in both rodents and humans (Gerber D et. al, 2021). We note similar results for NR2F2.

Table B.14: 1-Sided t-Test Comparison for Mean Square Error Values for Schwann Cell Applications (NetREm versus Benchmarks)

1-Sided t Test Comparison: Test Mean Square Error (MSE) Values for Schwann Cells SCs (Before Filtering)											
Alternative Hypothesis: Group 1 has a significantly lesser Test MSE than Group 2 does.											
SCs Type	Group 1	Group 2	Test statistic *t*	Degrees of freedom (df)	p-value	Mean		95% Confidence Bounds		# of Target Genes (ALL)	
						Group 1	Group 2	Low	Upper	Group 1	Group 2
mSC	NetREm	Ridge CV	-50.01	24,762.28	0	1.1761	2.211	-inf	-1.001	12,396	12,405
		Linear Regression	-81.59	23,389.55	0	1.1761	3.0820	-inf	-1.8675		12,409
		ElasticNet CV	7.8038	15,954.37	1	1.1761	1.0469	-inf	0.1565		3,614
		Lasso CV	7.462	15,345.50	1	1.1761	1.0513	-inf	0.15229		3,252
nmSC		Ridge CV	-7.242	26,904.00	2.27E-13	1.0957	1.1936	-inf	-0.0757	13,454	13,452
		Linear Regression	-7.338	26,903.00	1.12E-13	1.0957	1.1949	-inf	-0.0770		13,451
		ElasticNet CV	8.3306	16,382.44	1	1.0957	1.0113	-inf	0.10104		6,152
		LassoCV	8.485	16,288.37	1	1.0957	1.0099	-inf	0.10242		5,984
SCs Type	Group 1	Group 2	Test statistic *t*	Degrees of freedom (df)	p-value	Mean		95% Confidence Bounds		# of Target Genes (Common)	
						Group 1	Group 2	Low	Upper	Group 1	Group 2
mSC	NetREm	Ridge CV	-75.556	6,007.54	0	1.0432	2.1253	-inf	-1.0585	3,248	
		Linear Regression	-83.124	4,184.81	0	1.0432	3.0289	-inf	-1.9464		

		ElasticNet CV	-0.439	6,493.84	0.330495	1.0432	1.0485	-inf	0.01465	
		Lasso CV	-0.655	6,493.83	0.256247	1.0432	1.0512	-inf	0.01203	
nmSC		Ridge CV	-24.617	11,954.53	0	1.0218	1.1366	-inf	-0.10714	5,980
		Linear Regression	-24.911	11,954.44	0	1.0218	1.1380	-inf	-0.10853	
		ElasticNet CV	2.613	11,957.94	0.995513	1.0218	1.0097	-inf	0.019714	
		LassoCV	2.578	11,957.94	0.99502	1.0218	1.0099	-inf	0.019548	

The results of the 1-sided t-test comparison for Mean Square Error Values for Schwann Cells (SCs) being less than those predicted by the other 4 benchmark models (RidgeCV, Linear Regression, ElasticNetCV, LassoCV) at the 5% level. We utilize the original target genes (TGs) predicted by the models without performing any filtering based on test Mean Square Errors (MSEs). Please note that for each pairwise Welch test, we use the R software $t.test$ function default of $var.equal = False$.

Table B.15: Overlap in the top 500 Random Forest (RF) models for predicting neurodegenerative diseases across cell-types

Overlaps in # of top 500 genes based on RF classifier models for predicting neurodegenerative disease or not based on TF-TF coordination score changes $\bar{B}$ from Control to Alzheimer's disease (AD) Stages. This is a balanced classification task (undersampling class 0: True and/or False Negatives to equal class 1 links) based on the Contextual PPI database (PPID) (Kotlyar et. al 2022). For instance: Astrocytes (Astro) share 24 TGs with Oligodendrocytes (Oligo); both cross-talk (via cell-cell contact, cytokines, signal molecules) is core for neuroinflammation, glial development, neuron regeneration (Nutma et al. 2020).

	Pericytes	OPCs	Oligo	Microglia	In. Neuron	Ex. Neuron	Endo. BBB	Astrocytes
Pericytes		22	25	25	27	23	33	28
OPCs	22		26	21	35	26	25	25
Oligo	25	26		29	28	22	19	24
Microglia	25	21	29		26	16	29	21
In. Neuron	27	35	28	26		28	19	27
Ex. Neuron	23	26	22	16	28		19	27
Endo. BBB	33	25	19	29	19	19		15
Astrocytes	28	25	24	21	27	27	15	

Table B.16: Expanding NetREm to solve problems in biology and beyond

We present potential applications of NetREm to solve various problems in the field of bioinformatics and beyond. NetREm creates interpretable models that can be used to guide regression solutions, incorporate prior knowledge, and uncover relationships among the predictors. In addition to gene expression regression, NetREm can be extended to other emerging single-cell omics such as scATAC-seq to identify TF and chromatin interactions on open accessible regions at the cell-type level. By defining these components (predictors, response, input network, and goal), network-regularized regression can be effectively tailored to address specific bioinformatics challenges, leveraging prior knowledge to enhance predictive accuracy and biological interpretability. NetREm can be used for problems with continuous predictors (X) and response variable (y) where predictors (i.e. features) interact in functional subnetworks

(that can be provided as input prior knowledge). It predicts which predictor groups impact y and how they associate and coordinate in networks to influence the target y.

Application	Predictors (X)	Response Variable (y)	Input Network among Predictors	Goal
GRN Inference	Gene expression levels of transcription factors (TFs)	Expression levels of a target gene (TG)	Protein-protein interaction network (PPIN) or known co-regulatory interactions among TFs	Predict TG expression based on TF expression, identifying key TFs and their coordination.
Protein Function Prediction	Gene expression levels or protein abundances	Functional annotations or scores (e.g., enzymatic activity)	Protein-protein interaction network (PPIN)	Predict functions of uncharacterized proteins based on known interactions.
Disease Gene Identification	Genetic/genomic data (e.g., gene expression levels, SNP genotypes)	Disease phenotype presence/absence or severity score	Protein-protein interaction network (PPIN) or gene-disease association networks	Identify genes associated with diseases and how genetic variations influence disease risk.
Pathway Analysis	Gene expression levels or protein abundances	Pathway activation scores or pathway-specific gene expression levels	Pathway interaction network or gene-gene interaction network within pathways	Predict activation status of biological pathways and discover/refine pathway interactions.
Single-Cell Analysis	Single-cell RNA-seq gene expression levels	Cell type-specific gene expression profiles or differentiation states	PPI network or similarity network from single-cell regulatory network embeddings	Infer cell type-specific regulatory networks and model dynamic changes during differentiation.
Drug Response Prediction	Gene expression levels, protein abundances, or genetic variants	Drug response measures (e.g., sensitivity, resistance)	Drug-target interaction network or gene-drug interaction network	Predict individual responses to drugs based on molecular data for personalized medicine.
Epigenomics	Chromatin state features, histone modification levels, or DNA methylation	Gene expression levels or chromatin state activation scores	Chromatin interaction network or histone modification interaction network	Predict impact of chromatin states and histone modifications on gene regulation.

Integration of Multi-Omics Data	Combined features from multiple omics layers	Phenotypic traits, disease states, or molecular signatures	Integrated multi-omics network combining interactions from various omics layers	Build comprehensive models leveraging multi-omics data for complex traits and diseases.
Non-Coding SNP Impact Analysis	Genotype data for non-coding SNPs	Gene expression levels, chromatin accessibility, or regulatory activity	Regulatory network connecting SNPs to regulatory regions and target genes	Predict impact of non-coding SNPs on gene expression and regulatory activity.
Epistasis Analysis	Genotype data for pairs or groups of SNPs	Phenotypic traits or disease states	Epistatic interaction network encoding interactions between genetic variants	Identify and quantify significant epistatic interactions influencing complex traits/disease.
Polygenic Risk Score Calculation	Genotype data for multiple genetic variants	Disease phenotype or quantitative trait measurement	Network capturing interactions among genetic variants (SNP-SNP or gene-gene interaction network)	Calculate individual polygenic risk scores considering genetic variant interactions.

Table B.17: Metrics on TFs and TGs and samples across cell-types and conditions for fixed cell-type TFs (no prior GRN info)

Keeping the # of candidate Transcription Factors (TFs) constant (fixed) across all target genes (TGs): $N = \mathcal{N}$ for TGs that are not candidate TFs. For TGs that are also cell-type TFs, $N = \mathcal{N} - 1$ since we will remove that TF from the list of candidate TFs.

Cell-type and conditions	**# of TGs**	**# of candidate TFs $N = \mathcal{N}$** for $\mathcal{N}$ cell-type TFs (If the TG is a TF, then $N = \mathcal{N} - 1$): Maximum $\mathcal{N}$ = $\max(N)$)	**# of cells M** (if 100% of data is used for training, M = # of cells; else M = # of train cells)
Simulated Human Embryonic Stem Cells (hESCs) Case 1: $M > N$	1,250	$\mathcal{N} = 207$ cell-type TFs N = 206 to 207 cell-type TFs	(1,000; $M = 700$ train, 300 test)
Simulated hESCs Case 2: $M < N$			(100; $M = 70$ train, 30 test)
Human Hematopoietic stem cells (HSCs)	10,588	$\mathcal{N} = 178$ cell-type TFs N = 177 to 178 cell-type TFs	$M = 2{,}268$

Mouse Embryonic Stem cells (mESCs)	19,225	$\mathcal{N} = 195$ cell-type TFs N = 194 to 195 cell-type TFs	$M = 1{,}080$
Mouse Dendritic cells (mDCs)	9,087	$\mathcal{N} = 93$ cell-type TFs N = 92 to 93 cell-type TFs	$M = 1{,}211$
Naïve CD4-T cell (C0)	13,714 (PPBCs)	$\mathcal{N} = 1{,}029$ cell-type TFs N = 1,028 to 1,029 cell-type TFs	$M = 709$
CD14-Monocytes (C1)			$M = 480$
Memory CD4-T (C2)			$M = 429$
B cells (C3)			$M = 342$
CD8-T cell (C4)			$M = 316$
FCGR3A Monocyte (C5)			$M = 162$
Natural Killer cell (C6)			$M = 154$
Dendritic Cell (C7)			$M = 32$
PBMCs Platelet (C8)			$M = 14$
<u>Comparison with RTNduals:</u> Human Applications Filtering for TFs found in this study: (Göös et al. 2022) and in the gene expression dataset			
Schwann Cells GTEx Data (Eraslan et. al 2022) Pooled from 5 tissues	14,144	$\mathcal{N} = 88$ cell-type TFs N = 87 to 88 cell-type TFs	$M = 1{,}430$
Excitatory Neuron	15,693 (All Data) (Lake et al. 2018)	$\mathcal{N} = 76$ cell-type TFs N = 75 to 76 cell-type TFs	$M = 13{,}709$
Inhibitory Neuron		$\mathcal{N} = 76$ cell-type TFs N = 75 to 76 cell-type TFs	$M = 6{,}045$
Microglia		$\mathcal{N} = 97$ cell-type TFs N = 96 to 97 cell-type TFs	$M = 317$
Oligodendrocytes		$\mathcal{N} = 80$ cell-type TFs N = 79 to 80 cell-type TFs	$M = 2{,}657$
Control Excitatory Neurons	17,926 (Training Data) (Mathys et. al 2019)	$\mathcal{N} = 77$ cell-type TFs N = 76 to 77 cell-type TFs	$M = 11{,}969$
Control Inhibitory Neurons		$\mathcal{N} = 78$ cell-type TFs N = 77 to 78 cell-type TFs	$M = 3{,}378$
Control Microglia		$\mathcal{N} = 97$ cell-type TFs N = 96 to 97 cell-type TFs	$M = 676$
Control Pericytes		$\mathcal{N} = 98$ cell-type TFs N = 97 to 98 cell-type TFs	$M = 64$
Control Astrocytes		$\mathcal{N} = 93$ cell-type TFs N = 92 to 93 cell-type TFs	$M = 1{,}093$

Control Endothelial Blood Brain Barrier		$\mathcal{N} = 93$ cell-type TFs N = 92 to 93 cell-type TFs	$M = 43$
Control Oligodendrocyte Precursor Cells (OPCs)		$\mathcal{N} = 90$ cell-type TFs N = 89 to 90 cell-type TFs	$M = 936$
Control Oligodendrocytes		$\mathcal{N} = 82$ cell-type TFs N = 81 to 82 cell-type TFs	$M = 6{,}440$

Table B.18: Metrics on TFs and TGs and samples across cell-types and conditions for customized TG-specific cell-type TFs (with prior GRN info)

Using multi-omics data and prior gene regulatory network (GRN) knowledge to customize the N candidate TF predictors for each target gene (TG).

Cell-type and conditions	**# of TGs**	**# of candidate TFs N:** (aggregate metrics across TGs: Average $N = \bar{N}$, Median $N = med(N)$, Minimum $N = \min(N)$, Maximum $N = \max(N)$)	**# of cells** (70% training = M, 30% testing)
Astrocytes (AD and Control)	13,804	AD: $\{\bar{N} = 191.81; med(N) = 203;$ $\min(N) = 47;$ $\max(N) = 259\}$ Control: $\{\bar{N} = 193.12;$ $med(N) = 205;$ $\min(N) = 47;$ $\max(N) = 259\}$	AD: (1,830, M = 1,281 train, 549 test) Control: (1,562, M = 1,093 train, 469 test)
Endothelial Blood Brain Barrier (BBB) (AD and Control)	12,619	AD: $\{\bar{N} = 182.07; med(N) = 191;$ $\min(N) = 46;$ $\max(N) = 267\}$ Control: $\{\bar{N} = 166.22; med(N) = 173;$ $\min(N) = 49;$ $\max(N) = 243\}$	AD: (59, M = 41 train, 18 test) Control: (62, M = 19 train, 43 test)
Excitatory Neuron (AD and Control)	14,018	AD: $\{\bar{N} = 169.04;$ $med(N) = 184;$ $\min(N) = 41;$ $\max(N) = 219\}$ Control: $\{\bar{N} = 168.29; med(N) = 184;$ $\min(N) = 39;$ $\max(N) = 221\}$	AD: (17,878, M = 12,515 train, 5,363 test) Control: (17,098, M = 11,969 train, 5,129 test)
Inhibitory Neuron (AD and Control)	13,700	AD: $\{\bar{N} = 145.32; med(N) = 155;$ $\min(N) = 33;$ $\max(N) = 205\}$ Control: $\{\bar{N} = 148.92; med(N) = 159;$ $\min(N) = 34;$ $\max(N) = 209\}$	AD: (4,371, M = 3,060 train, 1,311 test) Control: (4,825, M = 3,378 train, 1,447 test)
Microglia (AD and Control)	12,490	AD: $\{\bar{N} = 179.99;$ $med(N) = 191;$ $\min(N) = 49;$ $\max(N) = 251\}$ Control: $\{\bar{N} = 182.62; med(N) = 195;$ $\min(N) = 52;$ $\max(N) = 252\}$	AD: (955, M = 668 train, 287 test)

			Control: (965, $M = 676$ train, 289 test)
Oligodendrocytes (AD and Control)	11,859	AD: $\{\bar{N} = 177.17; med(N) = 189; \min(N) = 54; \max(N) = 241\}$ Control: $\{\bar{N} = 173.95; med(N) = 185; \min(N) = 52; \max(N) = 234\}$	AD: (9,035, $M = 6{,}324$ train, 2,711 test) Control: (9,200, $M = 6{,}440$ train, 2,760 test)
Oligodendrocyte Precursor Cells (OPCs) (AD and Control)	12,641	AD: $\{\bar{N} = 172.45; med(N) = 184; \min(N) = 50; \max(N) = 234\}$ Control: $\{\bar{N} = 176.25; med(N) = 188; \min(N) = 50; \max(N) = 237\}$	AD: (1,290, $M = 903$ train, 387 test) Control: (1,337, $M = 936$ train, 401 test)
Pericytes (AD and Control)	14,206	AD: $\{\bar{N} = 210.55; med(N) = 230; \min(N) = 62; \max(N) = 278\}$ Control: $\{\bar{N} = 206.92; med(N) = 227; \min(N) = 55; \max(N) = 272\}$	AD: (76, $M = 64$ train, 27 test) Control: (91, $M = 53$ train, 23 test)
Myelinating Schwann Cells (mSCs)		$\{\bar{N} = 128.49; med(N) = 140; \min(N) = 28; \max(N) = 197\}$	(319, $M = 223$ train, 93 test)
Non-myelinating Schwann Cells (nmSCs)		$\{\bar{N} = 136.85; med(N) = 150; \min(N) = 27; \max(N) = 201\}$	(2,468, $M = 1{,}727$ train, 741 test)

Table B.19 Availability of Data and Materials

Data Resource	Access Information:
Mapping of cis-Candidate Regulatory Elements (cCREs) for hg38 human reference genome coordinates	http://catlas.org/catlas_downloads/humantissues/cCRE_hg38.tsv.gz.
Activity by Contact (ABC) scores for interacting cis-candidate REs (cCREs)	http://catlas.org/catlas_downloads/humantissues/ABC_scores/
Raw single-cell scATAC-seq peaks	http://catlas.org/catlas_downloads/humantissues/Peaks/RAW/
Single-cell DRG expression data	GSE169301
Rat SOX10 PNS binding tracks	GSE64703 (Lopez-Anido et. al 2015)
UCSC Genome Browser Session for SOX10 peaks	genome.ucsc.edu/s/saniyaKhullar/updated_rn5
EGR2 data	GSE35132
Raw depositions for H3K27ac ChIP-seq data and Cut&Run data (YY1, RXRG, TEAD1)	GSE247955

STAT1 data	GSE211338
Expression data for 4 mESC samples	GSE108222
Ground truth signed TF-TG hESC network	File 5 in (Sharov et al. 2022)
(Zhang et al. 2023) has HSC expression data	zenodo.org/records/7879228 (Buenrostro_Hematopoiesis.tar.gz)
(Zhang et al. 2023) has gold standard GRNs for HSCs and mESCs	zenodo.org/records/7879228 (scMTNI_sourcedata.tar.gz)
mDCs gene expression data	zenodo.org/records/5909090: normalized GSE48968 (expression_data.zip)
mDCs gold standards	zenodo.org/records/5909090: gold standards (gold_standard_datasets.zip)
mDCs comparative cell-type-GRNs	zenodo.org/records/5909090: comparative cell-type-GRNs (normalized_inferred_networks.zip).
SERGIO tool	github.com/PayamDiba/SERGIO
Contextual PPID	iid.ophid.utoronto.ca/
Older V11 STRINGdb PPIN	version-11-0.string-db.org/
Current PPIN includes V12 PPIs	string-db.org/
Tibial Nerve v8 eQTLs	gtexportal.org/home/downloads/adult-gtex/qtl
CNS eQTLs (Bryois et al. 2022)	zenodo.org/records/7276971

§ B.4 Supplementary files and information

Please click the respective hyperlinks to navigate to the following Supplementary Files for NetREm, which are publicly available online.

File B1

Schwann Cell (SC) Transcription Factor – Target Gene interaction ($|coef| \geq 0.05$). [💻🖱Link to data]

File B2

SC Transcription Factor (TF) – TF coordination ($|\bar{B}| \geq 0.01$, $|coef| \geq 0.025$) [💻🖱Link to data].

File B3

Cell-type TF-TG networks across AD and Control ($|coef| \geq 0.05$, test mean square error (MSE) ≤ 1) [💻🖱Link to data].

File B4

Cell-Type TF-TF coordination across AD and Control ($|\bar{B}| \geq 0.01$ and test MSE ≤ 1) [💻🖱Link to data].

References

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Grüning BA, et al. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research* **46**: W537–W544.

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 S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J, et al. 2017. SCENIC: Single-cell regulatory network inference and clustering. *Nat Methods* **14**: 1083–1086.

Ali O, Farooq A, Yang M, Jin VX, Bjørås M, Wang J. 2022. abc4pwm: affinity based clustering for position weight matrices in applications of DNA sequence analysis. *BMC Bioinformatics* **23**: 83.

Allen Brain Atlas Data Portal. Microarray Data :: Allen Brain Atlas: Human Brain. https://human.brain-map.org/ (Accessed May 27, 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.

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.

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.

Bhuva et. al 2021. msigdb. *Bioconductor*. http://bioconductor.org/packages/msigdb/ (Accessed May 27, 2024).

Bryois J, Calini D, Macnair W, Foo L, Urich E, Ortmann W, Iglesias VA, Selvaraj S, Nutma E, Marzin M, et al. 2022. Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat Neurosci* **25**: 1104–1112.

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.

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).

Carlson et. al 2019. org.Hs.eg.db. *Bioconductor*. http://bioconductor.org/packages/org.Hs.eg.db/ (Accessed September 26, 2023).

Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, et al. 2022. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **50**: D165–D173.

Cermenati G, Audano M, Giatti S, Carozzi V, Porretta-Serapiglia C, Pettinato E, Ferri C, D'Antonio M, De Fabiani E, Crestani M, et al. 2015. Lack of Sterol Regulatory Element Binding Factor-1c Imposes Glial Fatty Acid Utilization Leading to Peripheral Neuropathy. *Cell Metabolism* **21**: 571–583.

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.

Chen J, Xiao L, Rao JN, Zou T, Liu L, Bellavance E, Gorospe M, Wang J-Y. 2008. JunD Represses Transcription and Translation of the Tight Junction Protein Zona Occludens-1 Modulating Intestinal Epithelial Barrier Function. *Mol Biol Cell* **19**: 3701–3712.

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.

Cohen E. 2023. Node2Vec. https://github.com/eliorc/node2vec (Accessed September 3, 2023).

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.

Datta RR, Rister J. 2022. The power of the (imperfect) palindrome: Sequence-specific roles of palindromic motifs in gene regulation. *BioEssays* **44**: 2100191.

Dibaeinia P. 2024. PayamDiba/SERGIO. https://github.com/PayamDiba/SERGIO (Accessed February 17, 2024).

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

Drew K, Wallingford JB, Marcotte EM. 2021. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Molecular Systems Biology* **17**: e10016.

Durinck S, Spellman PT, Birney E, Huber W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**: 1184–1191.

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).

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).

Fogarty et. al 2020. SOX10-regulated promoter use defines isoform-specific gene expression in Schwann cells | BMC Genomics | Full Text. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-06963-7 (Accessed September 7, 2023).

Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, et al. 2020. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **48**: D87–D92.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**: R80.

Gerber et. al 2021. GEO Accession viewer. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE137870 (Accessed September 29, 2023a).

Gerber et. al 2021. Transcriptional profiling of mouse peripheral nerves to the single-cell level to build a sciatic nerve ATlas (SNAT) | eLife. https://elifesciences.org/articles/58591 (Accessed September 29, 2023b).

Giambartolomei et. al, 2014. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics | PLOS Genetics. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004383 (Accessed November 16, 2023).

Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Ruepp A. 2019. CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Research* **47**: D559–D563.

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.

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. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5048072/ (Accessed May 29, 2021).

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

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.

Gutiérrez-Sacristán A, Hernández-Ferrer C, González JR, Furlong LI. 2017. psygenet2r: a R/Bioconductor package for the analysis of psychiatric disease genes. *Bioinformatics* **33**: 4004–4006. https://doi.org/10.1093/bioinformatics/btx506 (Accessed May 29, 2021).

Hastie T. 2020. Ridge Regularization: An Essential Concept in Data Science. *Technometrics* **62**: 426–433.

He et. al, 2010. Yy1 as a molecular link between neuregulin and transcriptional modulation of peripheral myelination | Nature Neuroscience. https://www.nature.com/articles/nn.2686 (Accessed September 7, 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).

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).

Jain A, Tuteja G. 2021. *TissueEnrich: Tissue-specific gene enrichment analysis*. Bioconductor version: Release (3.12) https://bioconductor.org/packages/TissueEnrich/ (Accessed April 24, 2021).

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.

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.

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.

Karolchik D, Hinrichs AS, Kent WJ. 2009. The UCSC Genome Browser. *Curr Protoc Bioinformatics* **CHAPTER**: Unit1.4.

Kelner J, Koehler F, Meka R, Rohatgi D. 2021. On the Power of Preconditioning in Sparse Linear Regression. http://arxiv.org/abs/2106.09207 (Accessed March 9, 2024).

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler and D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**: 996–1006.

Klee S, Stamps MT. 2022. Eigenvalues of Graph Laplacians Via Rank-One Perturbations. *The Quarterly Journal of Mathematics* **73**: 609–616.

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).

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.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559. https://doi.org/10.1186/1471-2105-9-559 (Accessed April 24, 2021).

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**: R25.

Langmead et. al 2012. Fast gapped-read alignment with Bowtie 2 | Nature Methods. https://www.nature.com/articles/nmeth.1923 (Accessed September 26, 2023).

Lawrence et. al 2013. Software for Computing and Annotating Genomic Ranges | PLOS Computational Biology. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003118 (Accessed March 12, 2024).

Lawrence M, Gentleman R, Carey V. 2009. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**: 1841–1842.

Lê Cao K-A, Boitard S, Besse P. 2011. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* **12**: 253.

Le et. al 2005. Nab proteins are essential for peripheral nervous system myelination | Nature Neuroscience. https://www.nature.com/articles/nn1490 (Accessed September 7, 2023).

Lemaître G, Nogueira F, Aridas CK. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* **18**: 1–5.

Lerdrup et. al 2016. An interactive environment for agile analysis and visualization of ChIP-sequencing data | Nature Structural & Molecular Biology. https://www.nature.com/articles/nsmb.3180 (Accessed September 26, 2023).

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 et. al 2022. WashU Epigenome Browser update 2022 | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/50/W1/W774/6567479?login=false (Accessed May 27, 2024).

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, et al. 2012. Extensive Promoter-centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell* **148**: 84–98.

Liu T. 2014. Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells. *Methods Mol Biol* **1150**: 81–95.

Liu T, Xia L, Yao Y, Yan C, Fan Y, Gajendran B, Yang J, Li Y-J, Chen J, Filmus J, et al. 2019. Identification of diterpenoid compounds that interfere with Fli-1 DNA binding to suppress leukemogenesis. *Cell Death Dis* **10**: 1–11.

Lopez-Anido C, Sun G, Koenning M, Srinivasan R, Hung HA, Emery B, Keles S, Svaren J. 2015. Differential Sox10 genomic occupancy in myelinating glia. *Glia* **63**: 1897–1914.

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

Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, Brignall R, Cafarelli T, Campos-Laborie FJ, Charloteaux B, et al. 2020. A reference map of the human binary protein interactome. *Nature* **580**: 402–408.

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

Luo W, Brouwer C. 2013b. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29**: 1830–1831. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt285 (Accessed May 29, 2021).

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).

Machlab D, Burger L, Soneson C, Rijli FM, Schübeler D, Stadler MB. 2022. monaLisa: an R/Bioconductor package for identifying regulatory motifs. *Bioinformatics* **38**: 2624–2625.

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).

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.

Meyer et. al 2008. minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information | BMC Bioinformatics | Full Text.

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-461 (Accessed January 28, 2024).

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.

Ndungu A, Payne A, Torres JM, Bunt M van de, McCarthy MI. 2020. A Multi-tissue Transcriptome Analysis of Human Metabolites Guides Interpretability of Associations Based on Multi-SNP Models for Gene Expression. *The American Journal of Human Genetics* **106**: 188–201.

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.

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.

Nott A, Holtman IR, Coufal NG, Schlachetzki JCM, Yu M, Hu R, Han CZ, Pena M, Xiao J, Wu Y, et al. 2019. Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science* **366**: 1134–1139.

Nutma E, van Gent D, Amor S, Peferoen LAN. 2020. Astrocyte and Oligodendrocyte Cross-Talk in the Central Nervous System. *Cells* **9**: 600.

Oh S, Abdelnabi J, Al-Dulaimi R, Aggarwal A, Ramos M, Davis S, Riester M, Waldron L. 2022. HGNChelper: identification and correction of invalid gene symbols for human and mouse. *F1000Res* **9**: 1493.

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.

Ouyang Z, Zhou Q, Wong WH. 2009. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences* **106**: 21521–21526.

Overmyer KA, Shishkova E, Miller IJ, Balnis J, Bernstein MN, Peters-Clarke TM, Meyer JG, Quan Q, Muehlbauer LK, Trujillo EA, et al. 2021. Large-Scale Multi-omic Analysis of COVID-19 Severity. *Cell Syst* **12**: 23-40.e7.

Pagès 2017. SNPlocs.Hsapiens.dbSNP144.GRCh37. *Bioconductor*. http://bioconductor.org/packages/SNPlocs.Hsapiens.dbSNP144.GRCh37/ (Accessed November 14, 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.

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.

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).

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

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.

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.

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, 2023).

Satija et. al 2015. Spatial reconstruction of single-cell gene expression data | Nature Biotechnology. https://www.nature.com/articles/nbt.3192 (Accessed April 10, 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).

Schep et. al 2023. motifmatchr.knit. https://bioconductor.org/packages/release/bioc/vignettes/motifmatchr/inst/doc/motifmatchr.html (Accessed August 25, 2023).

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).

Shannon P, Richards M. 2023. MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs. https://bioconductor.org/packages/MotifDb/ (Accessed August 25, 2023).

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.

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

Siddharthan R. 2010. Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix. *PLoS One* **5**: e9722.

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.

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.

Svaren J, Meijer D. 2008. The molecular machinery of myelin gene transcription in Schwann cells. *Glia* **56**: 1541–1551.

systemsbiology.org> SA <seth ament at, systemsbiologyg.org> PS <pshannon at, systemsbiology.org> MR <mrichard at. 2023. trena: Fit transcriptional regulatory networks using gene expression, priors, machine learning. https://bioconductor.org/packages/trena/ (Accessed November 7, 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.

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

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.

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).

Tsukanov AV, Mironova VV, Levitsky VG. 2022. Motif models proposing independent and interdependent impacts of nucleotides are related to high and low affinity transcription factor binding sites in Arabidopsis. *Front Plant Sci* **13**: 938545.

Vecchiarelli-Federico LM, Liu T, Yao Y, Gao Y, Li Y, Li Y-J, Ben-David Y. 2017. Fli-1 overexpression in erythroleukemic cells promotes erythroid de-differentiation while Spi-1/PU.1 exerts the opposite effect. *Int J Oncol* **51**: 456–466.

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, van Calster B, Steyerberg EW. 2019. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and Prognostic Research* **3**: 18.

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

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

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.

Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**: 15.

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).

Xu J, Zhang B, Cai J, Peng Q, Hu J, Askar P, Shangguan J, Su W, Zhu C, Sun H, et al. 2023. The transcription factor Stat-1 is essential for Schwann cell differentiation, myelination and myelin sheath regeneration. *Molecular Medicine* **29**: 79.

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 et. al 2014. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/31/4/608/2748221?login=false (Accessed May 27, 2024).

Yu G. *Chapter 1 Overview of semantic similarity analysis | Biomedical Knowledge Mining using GOSemSim and clusterProfiler*. https://yulab-smu.top/biomedical-knowledge-mining-book/semantic-similarity-overview.html (Accessed August 25, 2023a).

Yu G. *clusterProfiler: universal enrichment tool for functional and comparative study*. http://yulab-smu.top/clusterProfiler-book/ (Accessed April 24, 2021b).

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.

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 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 T, Zhang Z, Dong Q, Xiong J, Zhu B. 2020. Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biology* **21**: 45.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zhao Y, Li M-C, Konaté MM, Chen L, Das B, Karlovich C, Williams PM, Evrard YA, Doroshow JH, McShane LM. 2021. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine* **19**: 269.

Zhou et. al 2018. 2018. How to extract promoters positions. https://hzhou.scholar.harvard.edu/blog/how-extract-promoters-positions (Accessed August 25, 2023).

Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. 2019a. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**: 1523.

Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. 2019b. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**: 1523.

CS109A - Advanced Section 2: Regularization. https://harvard-iacs.github.io/2019-CS109A/a-section/a-section2/ (Accessed March 9, 2024a).

Decision Curve Analysis: A Novel Method for Evaluating Prediction Models - Andrew J. Vickers, Elena B. Elkin, 2006. https://journals.sagepub.com/doi/10.1177/0272989X06295361?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed (Accessed November 14, 2023b).

Estimating the size of the human interactome | PNAS. https://www.pnas.org/doi/10.1073/pnas.0708078105 (Accessed November 15, 2023c).

Identifying ChIP-seq enrichment using MACS | Nature Protocols. https://www.nature.com/articles/nprot.2012.101 (Accessed September 24, 2023d).

Index of /gbdb/hg19/liftOver. https://hgdownload.soe.ucsc.edu/gbdb/hg19/liftOver/ (Accessed September 27, 2023e).

Index of /goldenPath/rn5/liftOver. https://hgdownload.soe.ucsc.edu/goldenPath/rn5/liftOver/ (Accessed August 25, 2023f).

2021. *Liftover to Map Genomic Coordinates Between Genome Builds with Rtracklayer Bioconductor Package*. https://www.youtube.com/watch?v=942WLiy3eb4 (Accessed August 25, 2023).