Mapping the Molecular Landscape of ALS Through Integrative Omics and Network Pharmacology

A Thesis
submitted to
Indian Institute of Science Education and Research Pune
in partial fulfilment of the requirements for the
BS-MS Dual Degree Programme

by

Akash Dileep K

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

Date: March 2026

Under the guidance of

Supervisor: Dr. Rahul Kumar
Indian Institute of Technology Hyderabad

From June 2025 to Mar 2026

INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH PUNE

Certificate

This is to certify that this dissertation entitled **Mapping the Molecular Landscape of ALS Through Integrative Omics and Network Pharmacology** towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by **Akash Dileep K** at the Indian Institute of Technology Hyderabad, under the supervision of **Dr. Rahul Kumar** (Associate Professor, Department of Biotechnology) during the academic year 2025–2026.

Akash Dileep K
(Student)

Dr. Rahul Kumar
(Supervisor)

Declaration

I, hereby declare that the matter embodied in the report titled “Mapping the Molecular Landscape of ALS Through Integrative Omics and Network Pharmacology” is the results of the investigations carried out by me at the Indian Institute of Technology, Hyderabad from the period June 2025 to March 2026 under the supervision of Dr. Rahul Kumar and the same has not been submitted elsewhere for any other degree.

Akash Dileep K

AKASH DILEEP K
20211157
BS-MS
IISER PUNE

Contents

List of Figures

Abstract

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disorder characterised by progressive motor neuron loss and marked molecular heterogeneity. This thesis presents a two-part computational multi-omic framework utilising the Answer ALS cohort and publicly available single-cell RNA sequencing data to identify early-stage biomarkers and nominate cell-type-specific therapeutic candidates through network-guided drug repurposing.

In Part I, Weighted Gene Co-expression Network Analysis (WGCNA) of bulk RNA-seq and ATAC-seq data from the Answer ALS cohort revealed that biological sex is the dominant source of transcriptomic variance, completely masking disease-associated signals in unsupervised global analyses. Sex-stratified network construction identified a male-specific co-expression module enriched for non-coding RNAs at the DLK1–DIO3 locus, from which LASSO logistic regression distilled a seven-gene diagnostic signature with a cross-validated AUC of 0.719. Consensus clustering of chromatin accessibility data identified a spurious two-cluster structure driven by Y-chromosomal mapping artifacts, which resolved upon peak removal. Differential correlation analysis of paired ATAC-seq and RNA-seq data further revealed a significant regulatory inversion at the EIF5 locus, implicating dysregulated RAN translation as a potential upstream amplifier of C9orf72-associated pathology.

In Part II, publicly available single-cell RNA sequencing data from TDP-43 ALS cortex was used to construct cell-type-specific gene regulatory networks. Network proximity-based drug repurposing identified significant candidates in the microglial and oligodendrocyte modules, while the astrocyte and excitatory neuron networks lacked sufficient topological density to nominate candidates. In microglia, the top-ranked compounds converged on GABA-A receptor subunits, with additional pharmacologically distinct candidates including venlafaxine, nortriptyline, pregabalin, and iloperidone. In oligodendrocytes, the dominant signal was dopamine D2 receptor antagonism, with ropinirole, risperidone, ziprasidone, and progesterone representing the highest-priority candidates.

Together, these findings demonstrate that rigorous confounder correction, sex-stratified analyses, and cell-type-specific network frameworks are essential for extracting biologically meaningful signals from population-scale ALS omics data.

Acknowledgement

I would like to sincerely thank my supervisor, Dr Rahul Kumar at the Indian Institute of Technology Hyderabad, for allowing me to undertake this research project and for his guidance, encouragement, and insightful discussions throughout my tenure. His mentorship allowed me the freedom to explore ideas independently while providing the direction needed to keep the work grounded in scientific rigour. I am also grateful to Dr Leelavati Narlikar at the Indian Institute of Science Education and Research Pune for serving as an expert member of my thesis advisory committee and for her thoughtful evaluation and constructive suggestions, which helped improve the clarity of this work.

I would also like to appreciate the members of my lab for the many discussions, shared ideas, and the supportive research environment that made this project both intellectually engaging and enjoyable. I am thankful to the Indian Institute of Technology Hyderabad for providing the research infrastructure and facilities that enabled this work. I would also like to acknowledge the Indian Institute of Science Education and Research Pune, where I am pursuing my BS–MS degree, for the academic training and environment that enabled me to undertake this research

Contributions

Contributor name	Contributor role
Akash Dileep K	Conceptualization Ideas
Akash Dileep K	Methodology
Akash Dileep K	Software
-	Validation
Akash Dileep K	Formal analysis
Akash Dileep K	Investigation
Dr. Rahul Kumar	Resources
Akash Dileep K	Data Curation
Akash Dileep K	Writing - original draft preparation
-	Writing - review and editing
Akash Dileep K	Visualization
Dr. Rahul Kumar	Supervision
Dr. Rahul Kumar	Project administration
-	Funding acquisition

This contributor syntax is based on the Journal of Cell Science CRediT Taxonomy[1]

Use of Generative Artificial Intelligence

Generative and agentic AI systems, including Claude Sonnet 4.6, and Grammarly Generative AI, were used during the preparation of this thesis to support understanding of technical concepts, assist with debugging, help structure parts of the literature review, and improve language quality, including tone, grammar, and readability.

These tools were used only as assistive instruments and not as authoritative scientific sources. All substantive intellectual contributions, including research framing, methodological design, implementation, analysis, interpretation, and conclusions, are the author's own. Any material produced with AI assistance was reviewed, verified, and revised by the author before inclusion. Full responsibility for the contents of this thesis rests with the author.

[1] https://journals.biologists.com/jcs/pages/author-contributions

Chapter 1

Introduction

Amyotrophic lateral Sclerosis (ALS) is a rapidly progressing neurodegenerative disorder. The condition was identified in 1869 by Jean-Martin Charcot, a French neurologist. ALS is characterized by the degeneration of lower and upper motor neurons, particularly in the brain stem and spinal cord. This leads to a series of clinical symptoms, including increasing muscle weakness, stiffness, and widespread muscle wasting. Irrespective of the site of onset, this condition culminates in full paralysis, often leading to respiratory failure and death, usually occurring within two to five years after symptoms first appear. Most cases of the disease (approximately 90–95%) occur sporadically with no clear family history, while the remaining 5–10% of cases are classified as familial ALS. Despite decades of research identifying genetic variants such as mutations in *SOD1*, *TARDBP*, *FUS*, and the *C9orf72* hexanucleotide repeat expansion, the precise mechanisms driving the selective vulnerability of motor neurons in sporadic cases remain incompletely understood.[1] Currently, there is no cure for ALS. Approved pharmacological treatments, such as riluzole and edaravone, are primarily disease-modifying therapies that offer only minor extensions to survival. This is attributed to our lack of understanding of the molecular mechanisms underlying the disease.(Masrori and Van Damme, 2020)

1.1 Disease Modeling: From Post-Mortem Tissue to iPSC Technology

A primary bottleneck in ALS research has been the difficulty of accurately modeling the human central nervous system. Historically, researchers have relied heavily on post-mortem human brain and spinal cord tissues. While these samples provide direct insight into the human pathology, they present severe limitations for understanding the origins of the disease. Post-mortem tissues represent the absolute end-stage of the neurodegenerative process. By the time these samples are collected, the vast majority

of vulnerable motor neurons have already undergone apoptosis, and the remaining tissue is heavily infiltrated by reactive astrocytes and microglia, creating a state of extensive gliosis that masks the initial molecular triggers of the disease. Consequently, analyzing post-mortem spinal cord tissue makes it exceptionally difficult to distinguish between the primary drivers of motor neuron death and the secondary inflammatory responses resulting from tissue degradation.

To bypass the limitations of end-stage tissue, the field has shifted to human induced pluripotent stem cell (iPSC) technology. By reprogramming somatic cells such as fibroblasts or peripheral blood mononuclear cells derived directly from ALS patients, researchers can generate patient-specific pluripotent lines. These lines can then be directed to differentiate into spinal motor neurons *in vitro*. This approach allows the researchers to model early-stage disease pathology prior to the onset of widespread cell death.

The Answer ALS consortium is one such initiative that has expanded the scale of iPSC-based research. This program was established to generate multi-omic data, including whole-genome sequencing, transcriptomics, and epigenomics from over 1,000 patient-derived iPSC motor neuron lines.(Baxi et al., 2022) The primary objective of such large-scale profiling is to identify reliable early-stage biomarkers and define clinical-molecular subtypes within the highly heterogeneous ALS cases.

However, iPSC-derived data from the Answer ALS consortium introduces its own set of methodological challenges. Workman et al (2023) highlighted that the transcriptomic data are particularly driven by the expression of glial cells, which arises from the variability resulting from the iPSC differentiation protocol. The tissue of origin also introduces variability. Furthermore, biological sex emerged as a dominant source of variance, with male-derived lines generating higher proportions of mature motor neurons and exhibiting distinct stress-related metabolic profiles compared to female lines(Workman et al., 2023)

These challenges are not limited to transcriptomics. In an analysis of the epigenomic landscape of the Answer ALS cohort, Tsitkov et al. (2024) performed ATAC-seq on motor neurons derived from 380 ALS patients and 80 controls.(Tsitkov et al., 2024) The initial analysis revealed that chromatin accessibility was dictated by technical and biological confounders, including the patient's biological sex, genetic ancestry, and sequencing variance. It was only after applying statistical corrections for these covariates that a modest ALS male-associated signal was isolated. Importantly, the researchers showed that after correcting the data, epigenomic signatures in early stages could accurately predict the long-term clinical progression of the disease, achieving precision similar to well-established clinical biomarkers(Tsitkov et al., 2024)

Therefore, the first aim of this thesis is to explore the early-stage regulatory landscape of ALS *in vitro*. By applying statistical frameworks that have not previously been

explored in this context, we aim to isolate robust early-stage biomarkers and investigate the existence of molecular subtypes from the paired ATAC and RNA-sequencing data.

1.2 Drug Repurposing

The ultimate goal of identifying disease mechanisms and biomarkers is to translate these findings into viable therapeutics. However, traditional *de novo* drug discovery is an extraordinarily inefficient process. The standard pipeline involving target identification, compound synthesis, preclinical animal testing, and multiple phases of human clinical trials typically requires 10 to 17 years to complete. This is further exacerbated by the financial burden of the process, with the average cost of bringing a single new molecular entity to market estimated to be between $2 billion and $3 billion.(DiMasi et al., 2016)

Furthermore, the probability of success is notably low; historically, only about 10% of drugs entering Phase I clinical trials receive regulatory approval. In the context of neurodegenerative diseases like ALS, the attrition rate is even higher. These drugs fail in late-stage clinical trials primarily due to unforeseen toxicity, poor blood-brain barrier permeability, or a lack of efficacy when moving from animal models to heterogeneous human populations. Given that the median life expectancy for an ALS patient is under five years, this highly failure-prone developmental timeline is inadequate.

Drug Repurposing serves as a faster alternative for drug development. This involves using pre-existing clinically approved or investigational drugs that fall outside their original medical scope. Because repurposed candidates have already undergone rigorous Phase I clinical testing to establish basic human safety, pharmacokinetics, and tolerability profiles, the risk of late-stage clinical failure due to unexpected adverse events is significantly reduced. By bypassing the initial stages of basic safety profiling, drug repurposing reduces the overall developmental timeline to roughly 3 to 12 years and capital investment to an average of $300 million. The overall success rate for repurposed drugs reaching approval is estimated to be near 30%, substantially higher than *de novo* efforts. For a rapidly progressing disease like ALS, drug repurposing allows for faster adaptation of treatment strategies based on emerging data and insights.

1.3 Computational Frameworks for Drug Repurposing

The recent growth of biological databases has transformed drug repurposing from clinical observations into systematic, data-driven approaches using computational framework-

based methods. These computational strategies can broadly be categorized into expression-based, structure-based, and network-based approaches, each with its own set of distinct advantages and limitations.

1.3.1 Expression-Based Repurposing

Expression-based drug repurposing relies on the principle of transcriptomic signature matching. This method involves extracting a distinct gene expression profile (the up-down differentially expressed disease signature) from patient tissues and comparing it against large reference databases of drug-induced transcriptomic changes. The core assumption is that if a specific drug induces gene expression changes that are perfectly inversely correlated with the disease signature, it serves as a viable therapeutic candidate.

The most prominent platforms used for this approach are the CMap(Connectivity Map) and the LINCS(Library of Integrated Network-Based Cellular Signatures) Database. These databases contain thousands of transcriptomic profiles generated by treating cultured cell lines with a vast array of small molecules. However, the utility of these platforms in neurodegenerative research is severely limited due to their biological composition. The reference profiles in CMap and LINCS were generated almost entirely using immortalized human cancer cell lines, such as MCF7 (breast cancer) or PC3 (prostate cancer)(Lamb et al., 2006)(Subramanian et al., 2017). As a result, neuronal cell lines are underrepresented, limiting the relevance of these resources for neurodegenerative studies.

Cancer cell lines are known for their high proliferation rates, distinct metabolism, and drastically different epigenetic landscapes when compared to terminally differentiated, post-mitotic human neurons. A small molecule that effectively downregulates an apoptotic pathway in a breast cancer cell line may yield entirely divergent transcriptional effects in a human spinal motor neuron. Moreover, using standardized cell lines often overlooks crucial patient-specific genetic contexts, such as polygenic risk scores or specific repeat expansions, which are vital for understanding individual disease mechanisms.

As a result, while expression-based screening methods are highly efficient, the predictions they generate frequently fail to translate effectively to the nervous system. This disconnect underscores the fundamental biological differences between the reference data derived from cancer cell lines and the target tissue of interest, leading to challenges in accurately modeling neurodegenerative diseases.

1.3.2 Structure-Based Repurposing

Structure-based drug repurposing approaches, primarily molecular docking and ligand-binding simulations, utilize high-resolution 3D physical structures of the target proteins to predict the binding affinity of existing small molecules. These computational Simulations calculate the thermodynamic favorability of a drug fitting into the active allosteric sites of a specific protein, allowing researchers to screen thousands of compounds rapidly.

While highly precise, structure-based repurposing requires comprehensive prior knowledge of the target molecule. The researcher must know exactly which protein is driving the disease and must have access to its crystallized structural conformation. This presents a major obstacle in ALS research. Sporadic ALS lacks a single, definitive pathogenic driver. Furthermore, primary pathological hallmarks of the disease, such as the aggregation of the TDP-43 protein, are notoriously difficult to target directly. TDP-43 is an essential RNA-binding protein, and attempting to bind its functional domains with small molecules risks disrupting its critical physiological roles in healthy cells. Therefore, approaches that rely entirely on predefined, single-target structures struggle to address the systemic, polygenic nature of ALS.(Choudhury et al., 2022)

1.3.3 Network-Based Repurposing

Network-based drug repurposing offers a novel approach to validating drug targets, addressing the limitations associated with single-target dependencies and biases from individual cell lines. This methodology is grounded in the understanding that human diseases arise not from isolated gene mutations, but from disruptions within highly interconnected cellular subnetworks of the human interactome. In this context, diseases can be depicted as specific ”disease modules,” which are clusters of dysregulated genes, and any gene that resides within or immediately adjacent to this module is statistically probable to exert a modulatory effect on the disease. Network-based repurposing evaluates therapeutic potential by calculating the topological proximity between the known targets of an existing drug and the identified disease module within the interactome network.

The primary advantage of network-based approaches is that they are fundamentally unsupervised. They do not require *a priori* knowledge of a specific, defined structural target, nor do they rely on exact 1:1 transcriptomic reversal matching. Instead, they identify compounds that are capable of perturbing the broader physiological pathways associated with the disease.

1.4 The Necessity of Single-Cell Resolution in Network Drug Repurposing

Network-based approaches are effective for drug repurposing, but their accuracy relies heavily on the quality of the transcriptomic data used to define disease modules. Traditionally, these modules have been built using bulk RNA-sequencing data from post-mortem brain or spinal cord tissue in ALS.

Bulk RNA-sequencing averages gene expression levels from millions of cells, which poses a problem in heterogeneous diseases like ALS. This issue is particularly evident in ALS, where disease progression is influenced not only by motor neuron defects but also by the surrounding microenvironment. In healthy conditions, glial cells support the central nervous system. However, in ALS, microglia become hyper-reactive, triggering chronic inflammation(Geloso et al., 2017), astrocytes fail to maintain a stable extracellular environment(Yamanaka et al., 2008), and oligodendrocytes(Philips and Rothstein, 2013) struggle to support axon health. These distinct cell types show varied responses to the disease, and using a single averaged model merges these conflicting signals, leading to an inaccurate understanding of the pathology.

Modern single-cell RNA sequencing (scRNA-seq) techniques enable the profiling of individual cells, allowing researchers to obtain cell-specific expression data. This approach presents an opportunity to construct cell-specific regulatory networks, offering insights into potential druggable targets tailored to specific cellular contexts. Despite its potential, this application remains largely unexplored within the field of amyotrophic lateral Sclerosis (ALS). Consequently, our second thesis objective focuses on using single-cell RNA sequencing to develop a method for drug repurposing stratergies.

1.5 Problem Statement and Thesis Objectives

The large-scale generation of population-scale iPSC datasets, combined with the computational advancement of network biology and single-cell transcriptomics, offers a distinct opportunity to decode the architecture of Amyotrophic Lateral Sclerosis. To address these distinct analytical challenges, this thesis is structured into two complementary parts:

Part I focuses on the bulk transcriptomic and epigenomic alterations that characterize early-stage ALS pathology. Utilizing the extensive Answer ALS *in vitro* cohort, this section implements statistical frameworks that have largely remained unexplored in this context to explore the existence of latent molecular subtypes and identify reliable early-stage biomarkers that are active prior to the onset of widespread cellular apoptosis.

Part II involves using Single-Cell RNA sequencing data and network-based drug repurposing to infer cell-type-specific drug targets. This project seeks to develop a computationally validated framework for drug repurposing. This approach will specifically identify candidates designed to address the distinct failures of diverse cellular populations in amyotrophic lateral sclerosis (ALS).

Chapter 2

Methods

Part I

2.1 Weighted Gene Co-expression Network Analysis (WGCNA)

Weighted Gene Co-expression Network Analysis Langfelder and Horvath (2008) is a framework to infer correlation patterns among genes across large-scale datasets. It assumes a simple guilt-by-association principle, whereby genes with synchronous expression profiles are inferred to be functionally related and to share common pathways, etc. This approach reduces the dimensionality of the expression matrix into biologically interpretable modules.

2.1.1 Adjacency Matrix Calculation

WGCNA begins with an adjacency matrix $A = [a_{ij}]$ that calculates pairwise pearson correlations and stores it as connection weights between genes. Rather than employing a hard threshold where values above or below a given threshold are set to 0, and others are retained, WGCNA uses a soft thresholding system. Soft thresholding raises the absolute value of correlation to a power β

$$a_{ij} = \text{cor}(x_i, x_j)^{\beta} \tag{2.1}$$

This soft thresholding serves two purposes: first, it prioritizes stronger connections while retaining biologically meaningful weaker ones. Second, it helps approximate a scale-free network topology, a property of biological networks in which few genes are highly connected and many have fewer connections.

In practice, the soft-thresholding power β is selected such that the resulting network

approximates scale-free topology, typically by choosing the smallest β that yields a scale-free topology fit index (R^2) greater than ~ 0.8. (Zhang and Horvath, 2005)

2.1.2 Topological Overlap Measure (TOM)

While the adjacency matrix considers only the direct relationship between two genes, WGCNA employs an additional metric, the Topological Overlap Matrix (TOM), which incorporates a third neighbor into the pairwise correlation. This helps capture a higher-order network structure.

The topological overlap TOM_{ij} between genes i and j is defined as:

$$\text{TOM}_{ij} = \frac{l_{ij} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}} \tag{2.2}$$

Where:

- $l_{ij} = \sum_{k \neq i,j} a_{ik} a_{kj}$ represents the sum of the products of adjacencies through shared neighbors k.
- $k_i = \sum_j a_{ij}$ is the connectivity (degree) of gene i.
- The denominator $\min(k_i, k_j) + 1 - a_{ij}$ normalizes the measure to ensure $0 \leq \text{TOM}_{ij} \leq 1$.

2.1.3 Dissimilarity and Clustering

The TOM matrix is converted to a dissimilarity matrix, and hierarchical clustering is performed.

$$d_{ij}^{\text{TOM}} = 1 - \text{TOM}_{ij} \tag{2.3}$$

After clustering, the Dynamic Tree Cut algorithm is applied to the dendrogram. It identifies modules by adaptively cutting the tree based on its hierarchical structure, defining modules as distinct branches of the resulting dendrogram.

2.1.4 Module Eigengenes (ME)

Modules are represented by the Module Eigengene (ME), which is the first principal component of the module's expression matrix. The ME acts as a mathematically optimal summary of the "average" expression profile for all genes within a module across samples.

The calculation is performed via Singular Value Decomposition (SVD) on the standardized expression matrix $X^{(q)}$ of module q:

$$X^{(q)} = UDV^T \tag{2.4}$$

Where the first column of the matrix V corresponds to the module eigengene $E^{(q)}$.

2.1.5 Trait-associated Modules

The Biologically relevant modules are identified by correlating the module eigengenes with clinical traits (T). This relationship, termed Eigengene Significance, is quantified as:

$$\text{Significance} = \text{cor}(E^{(q)}, T) \tag{2.5}$$

A high absolute correlation indicates that the module's aggregate expression is significantly associated with the clinical phenotype of interest.

2.1.6 Hub Genes

Hub genes are highly connected, centrally positioned nodes that potentially drive the biological processes underlying a module. They are identified through two primary connectivity metrics:

- **Intramodular Connectivity** (k_{IM}): The sum of adjacencies within the module:

$$k_i = \sum_{j \in q,\, j \neq i} a_{ij} \tag{2.6}$$

- **Module Membership** (kME): The correlation between a gene's expression profile x_i and its module eigengene:

$$kME_i = \text{cor}(x_i, E^{(q)}) \tag{2.7}$$

Genes with $kME \approx 1$ are considered "intramodular hubs" and are prioritized as key biomarkers or therapeutic targets.

2.2 Biomarker Identification using Lasso

Lasso Penalty Regression was applied to the WGCNA-derived models to establish a minimal set of diagnostic genes for detecting ALS; we used the `glmnet` package for this

(Friedman et al., 2010). LASSO handles high-dimensional predictors by simultaneously regularizing and performing variable selection, producing a sparse, interpretable set of gene-level predictors.

This we get by minimizing the negative log-likelihood with an ℓ_1 penalty:

$$\hat{w} \;=\; \arg\min_{w} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right] \;+\; \lambda \|w\|_1 \right\}, \qquad p_i = \sigma(x_i^\top w), \tag{2.8}$$

The regularization parameter λ was selected via k-fold cross-validation using `cv.glmnet` function, with the optimal value chosen based on minimum cross-validated deviance (or the one-standard-error rule for increased sparsity). We retained genes with non-zero coefficients as candidate biomarkers, and the sign of each coefficient indicates how each gene is associated with disease status.. Model performance was evaluated using cross-validated area under the receiver operating characteristic curve (AUC–ROC).

2.3 Patient Subtyping

2.3.1 Latent Factor Extraction via nsNMF

Non-negative Matrix Factorization (NMF) is an unsupervised dimensionality-reduction method used to identify latent molecular subtypes and sparse gene expression programs from high-dimensional transcriptomic (Pascual-Montano et al., 2006) (Gaujoux and Seoighe, 2010). Standard Non-negative Matrix Factorization (NMF) decomposes a matrix into two lower-rank matrices: W and H. The matrix W represents genes versus latent factors, where the latent factors can be interpreted as biological programs(gene x Latent Factor Matrix). Meanwhile, H captures the contribution of each program across different patients(Latent Factor x Patient matrix). Here, we employed an extension of standard NMF that incorporates an intermediate square smoothing matrix S, such that the target matrix is approximated as $V \approx WSH$, where W is the basis matrix, H is the mixture coefficient matrix, and S is the smoothing matrix mathematically defined as

$$S = (1 - \theta)I + \left(\frac{\theta}{r}\right) J_r \tag{2.9}$$

where $\theta \in [0, 1]$ is a smoothing parameter, I is the $r \times r$ identity matrix, J_r is the $r \times r$ all-ones matrix, and r is the number of latent factors. When $\theta = 0$, S reduces to the identity matrix, recovering standard NMF; when $\theta = 1$, S becomes a uniform matrix, imposing maximal smoothing across factors. To identify latent transcriptomic

subtypes in this study, nsNMF was applied to the 5000 most variably accessible regions of the filtered gene expression matrix using the `NMF` R package (Gaujoux and Seoighe, 2010)with the smoothing parameter set to $\theta = 0.5$. This specific parameter provides a moderate smoothing effect, enforcing a controlled sparseness that enables the extraction of subtype-specific metagenes while preventing the model from becoming overly smooth. To ensure algorithmic stability and avoid local minima, the factorization was executed with `nrun = 1000` independent iterations utilizing random initial seeding. Global cluster stability was subsequently quantified using the cophenetic correlation coefficient derived from the resulting consensus matrix.

2.3.2 Consensus Clustering

Consensus clustering is an ensemble method that identifies stable molecular subtypes by aggregating results across multiple iterations of a base clustering algorithm. The framework was implemented using the R package `ConsensusClusterPlus`, which constructs a symmetrical $N \times N$ consensus matrix from repeated subsampling of samples and features (Wilkerson and Hayes, 2010). Each entry in this matrix represents the frequency with which a pair of samples is grouped, providing a quantitative measure of stability that is less sensitive to noise and the stochastic initialization of traditional algorithms like k-means. To apply the algorithm, we evaluate a range of potential cluster counts (k) by analyzing the Cumulative Distribution Function (CDF) of the consensus matrix. The CDF summarizes the distribution of consensus values across all sample pairs, where a flatter CDF plateau indicates more stable clustering. Once the optimal clusters are established, their clinical significance is validated by assessing associations with phenotypic variables. Categorical clinical data are analyzed using Fisher's Exact Test, while continuous, non-normally distributed variables are evaluated using the Wilcoxon Rank-Sum Test. We ran the ConsensusClusterPlus algorithm with 1,000 resampling iterations and a subsampling proportion of 80% for both samples and features with agglomerative hierarchical clustering as the base algorithm.

2.4 Transcription Factor Activity Prediction

High-throughput transcriptomic data is inherently high-dimensional and notoriously noisy, making steady-state mRNA levels a poor proxy for functional transcription factor (TF) activity. To overcome this, we use the PSIONIC(Osmanbeyoglu et al., 2019a) framework, a TF activity inference method that relies on a prior" set of regulatory relationships to gain deeper biological insight. This prior could be constructed from other databases or through ATAC datasets. In our context, we used paired ATAC-

seq chromatin accessibility data to construct a more informative, context-specific prior. The implementation of the methodology begins with the ChIPpeakAnno(Zhu et al., 2010) package, which is used to annotate open chromatin peaks to their nearest genes based on proximity. In parallel, each accessible chromatin peak is scanned against a curated TF binding motif reference database using FIMO(Grant et al., 2011) ($p < 10^{-5}$), which assigns each peak to one or more TF binding motifs; The maximum motif score across all peaks assigned to a gene is then used to populate the regulatory feature matrix X (genes $\times$ TF). To refine the resulting feature matrix and address redundancy in transcription factor binding profiles, we calculate a pairwise Jaccard Index among the TFs based on their binding sites and remove highly overlapping target sets. This preprocessing step ensures that the subsequent Group-ordered Multi-task Learning (GO-MTL)(Kumar and Daumé, 2012) framework receives a distinct and non-redundant set of regulatory signals as input for modeling. The framework further operates on Y (genes $\times$ patients), the normalized RNA-seq expression matrix, with the goal of learning W (TF $\times$ patients), a matrix of patient-specific TF activities, by solving the regression $Y \approx XW$ across all patients simultaneously. The core of the prediction framework employs the GO-MTL architecture to jointly model W as a factorization $W = LS$, where L (TF $\times$ K) contains K shared latent regulatory programs capturing co-regulated TF–gene relationships across the cohort, and S (K $\times$ patients) holds the sample-specific mixing coefficients that reconstruct each patient's regulatory profile as a linear combination of these K latent archetypes. This multi-task learning (MTL) approach allows the model to "borrow strength" across related patient samples, significantly improving the robustness of the inferred activities compared to single-task methods. Finally, to test the model's performance, we use Spearman's correlation between predicted and measured gene expression levels, providing a nonparametric validation of the model's accuracy in capturing transcriptional regulation

2.5 Regulatory Rewiring

To identify regulatory changes, we performed a differential correlation analysis of ATAC-seq peaks with their nearest genes. The ChIPpeakAnno R package(Zhu et al., 2010) was used to annotate the peaks to gene based on their proximity. We summed the peaks in the neighbourhood of each gene and computed the Spearman correlation between each gene's RNA expression and its matched ATAC accessibility score, separately for the ALS and control groups. To identify these rewired regulatory links, we compared the correlation coefficients from the two groups using Fisher's r-to-z transformation to calculate a z-difference score. Then a filter of q value $<$ 0.05 was applied to get the significant rewired peaks.

Part II

2.6 Single-Cell RNA Sequencing Data Processing

For the purposes of drug repurposing, we selected the pTDP-43-high samples and matched control samples from (Wang et al., 2025). Single-cell RNA sequencing data processing and quality control were performed using the `Scanpy` package in Python (Wolf et al., 2018). Raw count matrices were filtered to remove low-quality cells based on mitochondrial and ribosomal gene expression thresholds. The data underwent total-count normalization followed by log1p transformation. Principal Component Analysis (PCA) was utilized for initial dimensionality reduction, and batch effects across samples were corrected using the Harmony algorithm (Korsunsky et al., 2019). A k-nearest neighbours graph was constructed to compute the Uniform Manifold Approximation and Projection (UMAP) for two-dimensional visualization, and Leiden community detection was applied for cell clustering.

2.7 Cell Type Annotation

Cluster-specific marker genes were identified using the Wilcoxon rank-sum test, facilitating the manual annotation of distinct central nervous system cell types, including oligodendrocytes, microglia, astrocytes, and excitatory neurons. Cluster identities were determined by cross-referencing the identified marker genes against the PanglaoDB (Franzén et al., 2019) reference library, a curated database of cell type-specific marker genes. Given their established implications in ALS pathology, subsequent analyses were restricted to these four cell populations.

2.8 Pseudobulk Differential Expression Analysis

To identify disease-associated transcriptomic changes, cells from each annotated cluster were aggregated into pseudobulk profiles by summing raw counts across cells belonging to the same sample and condition. This pseudobulk approach accounts for within-sample correlation and enables the application of count-based statistical frameworks. Differential expression analysis between ALS and control pseudobulk profiles was performed using the `PyDESeq2` package (Muzellec et al., 2023), which implements the DESeq2 negative binomial model in Python. Genes passing a significance threshold of adjusted $p < 0.05$ and $|\log_2 \text{FC}| > 1$ were considered differentially expressed. The resulting sets of significantly dysregulated genes per cell type were subsequently used as disease seed genes for network propagation analyses.

2.9 Network Construction

For each cell type cluster in both the ALS and control conditions, we constructed a biological network using the SCENIC package. This yielded a hierarchical graph representing each transcription factor (TF) and its regulatory interactions with target genes. These graphs were used to calculate differential rewiring between conditions. However, for the subsequent drug repurposing step, TFs pose a challenge as they are largely not directly druggable. To address this, we computed pairwise cosine similarity between the gene–TF importance score vectors derived from SCENIC. This allowed us to convert the bidirected hierarchical graph into a undirected graph in which TFs are removed while their regulatory information is preserved

2.10 Calculating Differential Rewiring

Network topology was analyzed using the `NetworkX` library in Python to compute betweenness centrality for each transcription factor in both the ALS and control networks. A Disorder Influence (DI) score was calculated as the base-2 logarithm of the ratio between disease and control centrality.(Gupta et al., 2025)

$$\mathrm{DI} = \log_2\left(\frac{BC_{\text{disease}}}{BC_{\text{control}}}\right) \tag{2.10}$$

This quantified the regulatory influence gained or lost by each transcription factor during the disease state.

2.11 Candidate Drug Filtering for Neuro-Therapeutics

Drug-target interaction data were retrieved from DrugBank (Wishart et al., 2006) and DGIdb(Freshour et al., 2021), from which non-antineoplastic, FDA-approved drugs were selected. These candidates were subsequently cross-referenced with the Blood-Brain Barrier Database (B3DB)(Meng et al., 2021) to retain only compounds with confirmed BBB permeability,

2.12 Candidate Prioritization for Drug Targets

BBB-permeable drug-target interaction data were mapped onto the co-regulatory gene network. To assess the proximity of each drug's targets to the ALS disease module, differentially expressed genes from the annotated cell types were used as disease seeds. The proximity of a drug to the disease module was quantified using the mean

shortest path distance between the drug's target genes T and the disease gene set G, defined as:

$$d_{TG} = \frac{1}{|T|} \sum_{t \in T} \min_{g \in G} d(t, g) \tag{2.11}$$

where $d(t, g)$ denotes the shortest path length between target node t and disease gene g in the network. To establish statistical significance, a degree-preserving null model was constructed by binning network nodes within a degree difference of 10 and randomly sampling bin-matched nodes across 1,000 permutation iterations. The mean proximity distance across these iterations formed the null distribution, against which the observed drug-target proximity score was compared. A Z-score was then derived by standardizing the observed distance against this null distribution, and drugs were subsequently ranked based on this Z-score.

Chapter 3

Results

3.1 Global Network Analysis Reveals Confounding by Sex and other variables

To identify systems-level transcriptional changes associated with ALS, we initially constructed a weighted gene co-expression network using the entire AnswerALS cohort (N=1,100 samples). Soft-thresholding power selection was performed to achieve a scale-free topology fit index $(R^2 > 0.8)$.

However, module-trait association analysis revealed that the primary drivers of variance in the global network were not related to disease pathology. Several modules showed strong, significant correlations with biological confounders, most notably the biological sex of the participants. Crucially, no single module in the global analysis displayed a statistically significant association with the primary disease status (ALS vs. Control) after correction for multiple testing. This finding is consistent with our Principal Component Analysis (PCA), where PC1 was also strongly driven by sex differences rather than disease status.

Given the strong confounding effect of sex observed in the global network, we hypothesized that sexual dimorphism might be masking disease-relevant signals. To address this, we stratified the dataset and constructed independent co-expression networks for the male and female cohorts.

This stratification strategy proved effective. In the male-specific network, we identified a distinct co-expression module that was significantly correlated with ALS disease (Figure 3.1). This module (referred to hereafter as the green module) displayed a positive correlation with the ALS phenotype.

In contrast, the female-specific network failed to yield a module with a comparable association to disease status, suggesting that the transcriptional program captured in the male green module is highly specific to the male disease pathology. Consequently, we prioritized this male-specific module for downstream biomarker identification, using

its constituent genes as the input feature set for LASSO regression modeling

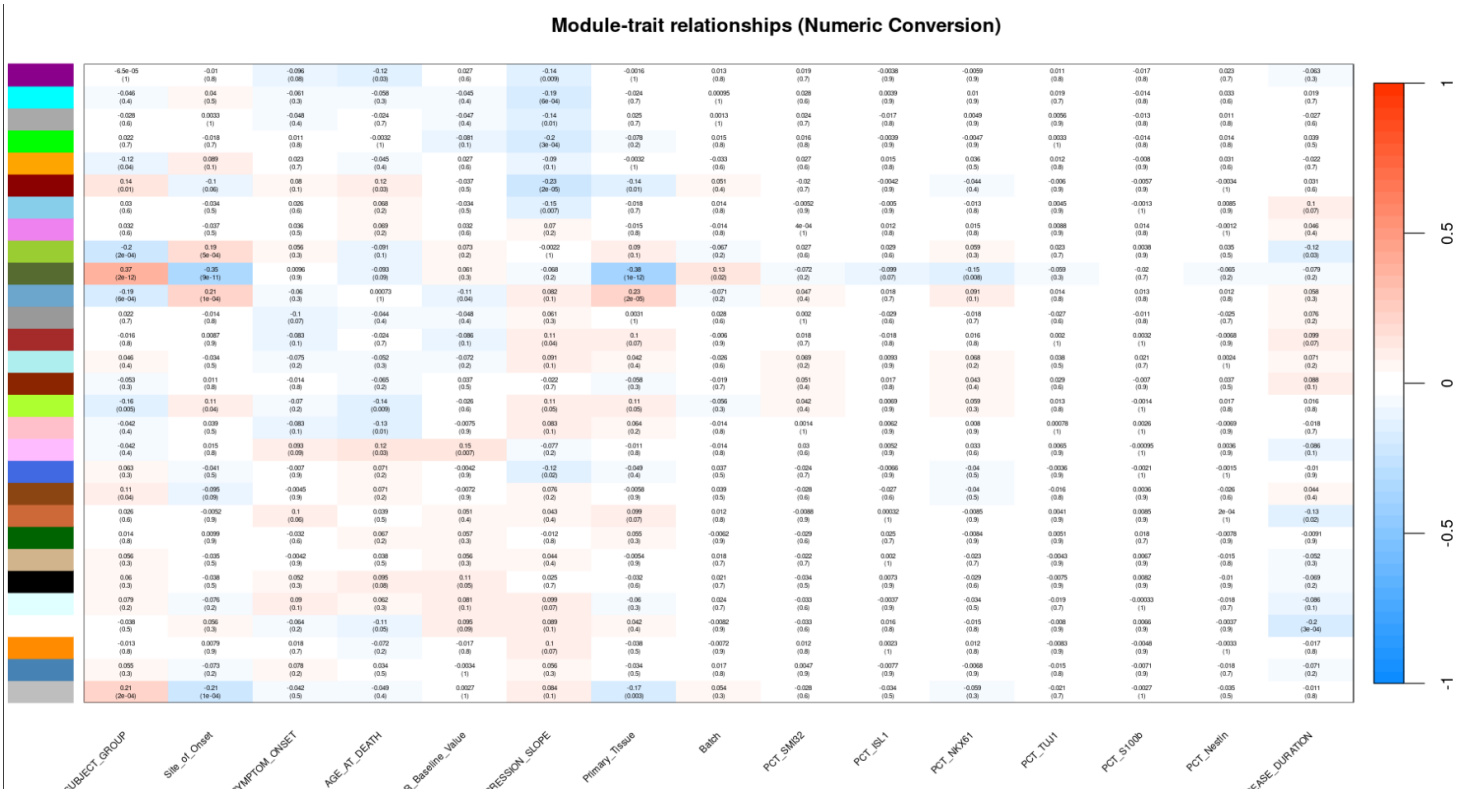


Figure 3.1: Heatmap of module-trait correlations. The Y-axis shows the co-expressed modules (by color) and the X-axis shows the traits. Each box contains the correlation value (top) and p-value (bottom)

3.2 Identification of a Male-Specific Diagnostic Gene Signature

We first characterized the composition of the significant male-specific co-expression module to understand the biological context of this disease driver. The module, comprising approximately 75 genes, displayed a striking enrichment for non-coding and regulatory genes rather than canonical protein-coding genes. Specifically, the module was heavily populated by the SNORD family of small nucleolar RNAs (e.g., SNORD113-6, SNORD114-3), which are known to guide chemical modifications of other RNAs. Additionally, we observed a dense cluster of microRNAs (e.g., MIR127, MIR370, MIR433) and Zinc Finger proteins (e.g., ZNF682, ZNF528-AS1), suggesting this module rep-

resents a coordinated regulatory network involved in transcriptional repression and post-transcriptional modification rather than a simple downstream effector pathway.

To distill this broad regulatory network into a clinically viable diagnostic panel, we applied LASSO (Least Absolute Shrinkage and Selection Operator) logistic regression to the module's gene set. This regularization method was selected to mitigate overfitting by penalizing complex models and shrinking the coefficients of less informative features to zero. From the initial 75-gene input, the algorithm converged on a minimal, sparse signature of 7 robust predictors (Figure 4.2)

The final LASSO-derived model achieved a cross-validated Area Under the Curve (AUC) of 0.719, indicating moderate but robust predictive power. The resulting 7-gene signature included FAM50B, LINC02301, and HSPA2 as positive predictors (associated with ALS cases), while genes such as ZNF682 were identified as negative predictors (associated with healthy controls).

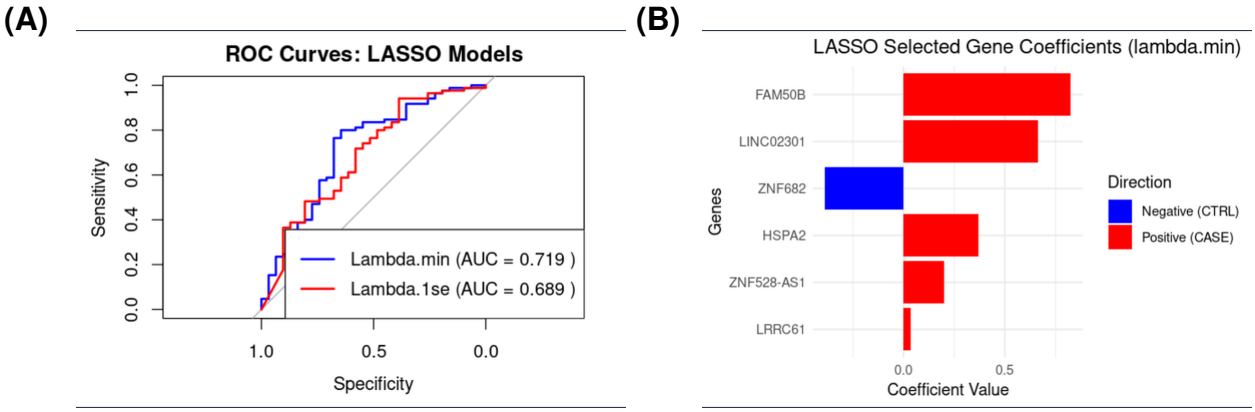


Figure 3.2: LASSO Model Performance and Gene Coefficients. (A) ROC curves evaluating the performance of the Lambda.min (AUC = 0.719) and Lambda.1se (AUC = 0.689) models. (B) Selected gene coefficients from the Lambda.min model.

3.3 Absence of Latent Transcriptomic Subtypes via NMF

To determine if the heterogeneous ALS cohort could be stratified into distinct molecular subtypes, we applied Non-negative Matrix Factorization (NMF) to the gene expression data. Despite utilizing an algorithm designed to enforce sparsity and enhance biological interpretability, the decomposition failed to yield stable or biologically relevant latent factors (Figure 4.3). The analysis did not identify consistent patient subgroups across a range of factorization ranks, as indicated by the lack of distinct patterns in the consensus matrix. Furthermore, the resulting basis vectors failed to capture coherent gene expression programs that could distinguish ALS

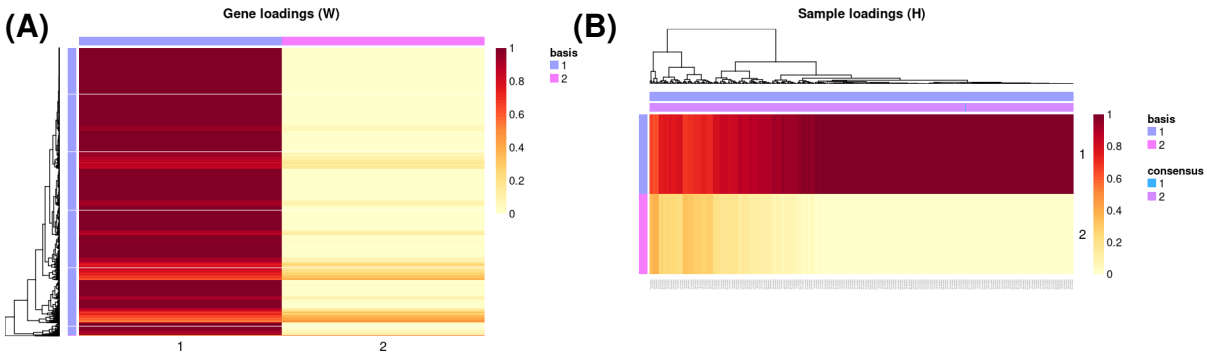


Figure 3.3: NMF Gene and Sample Loadings. (A) Gene loadings (W) plot showing gene contributions to latent factors 1 and 2. (B) Sample loadings (H) plot showing sample contributions to latent factors 1 and 2.

3.4 Identification of Technical Artifacts in Chromatin Accessibility Clustering

We initially applied consensus clustering to the RNA-seq dataset to identify robust patient subtypes based on transcriptional profiles. However, consistent with the NMF results, this analysis failed to yield any stable clusters or distinct patient subgroups.

In contrast, when we applied the same methodology to the ATAC-seq chromatin accessibility data, the algorithm initially identified two distinct and statistically stable clusters (k=2) (Figure 4.4). To investigate the biological drivers of this separation, we constructed a heatmap of the most variable accessible regions (VARs) across the cohort. Analysis of the top 100 most variable peaks revealed that the clustering was not driven by a broad genome-wide signature but rather by a specific, narrow locus of peaks.

Detailed annotation of these driving features identified them as Y-linked genomic regions. Inspection of the raw data showed that these peaks possessed a count value of exactly 1 in a specific subset of samples, effectively functioning as a binary technical variable rather than a continuous biological signal. To determine if these peaks had any functional consequencess on disease pathology, we integrated the matching transcriptomic data and investigated the expression of genes associated with these specific chromatin peaks. We found no significant alteration in gene expression of the genes associated with these peaks between the two identified clusters.

Consequently, we hypothesized that the clusters were a technical artifact driven by sex-specific mapping or low-level noise rather than disease biology. Upon removing these specific Y-linked peaks from the dataset and re-running the consensus clustering, the stable two-cluster structure ceased to exist. This confirmed that the initial

separation was an artifact and that no robust molecular subtypes were present in the chromatin accessibility data after quality control.

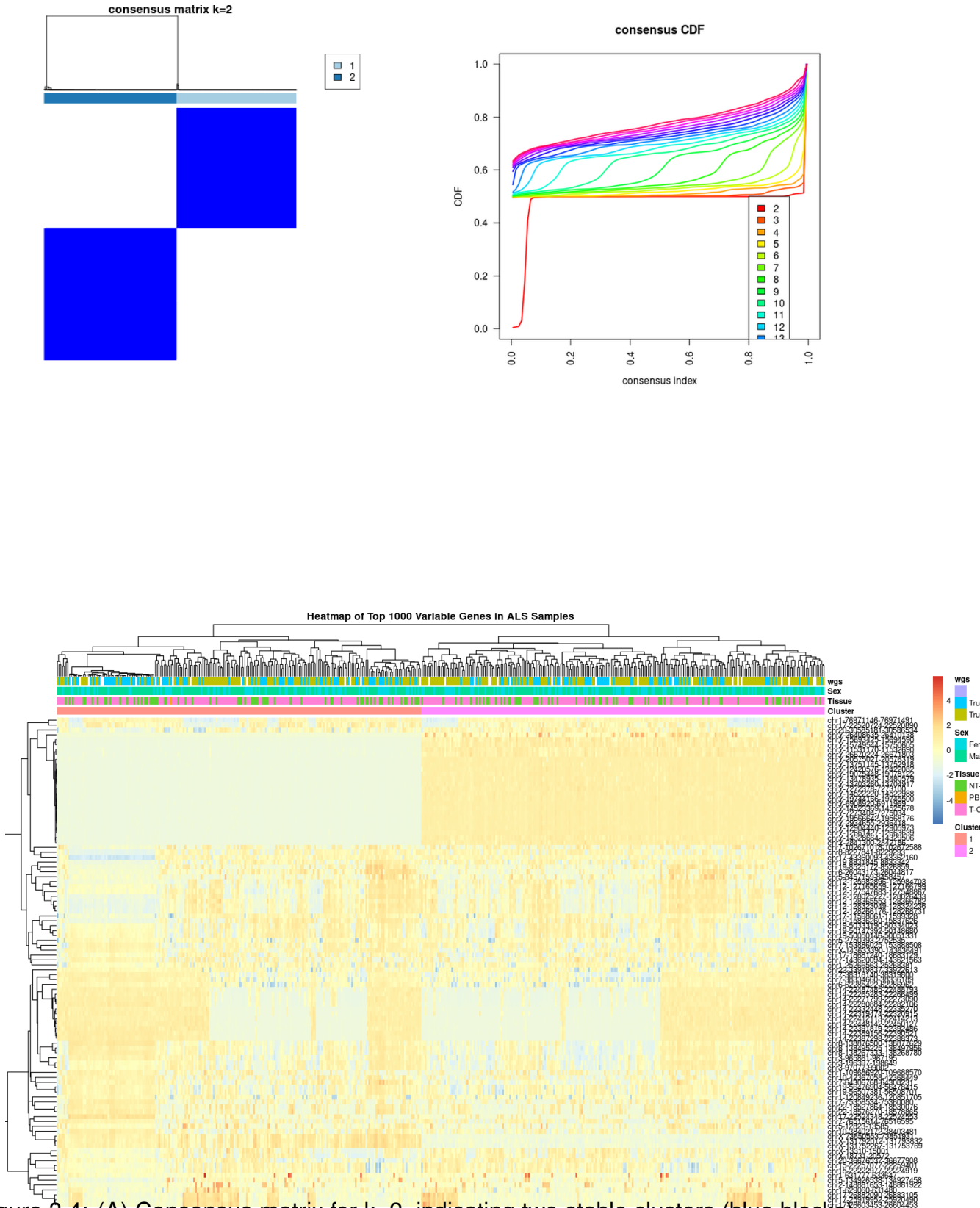


Figure 3.4: (A) Consensus matrix for k=2, indicating two stable clusters (blue blocks). (B) Consensus CDF plot for k=2 to 12. (C) Heatmap showing the expression values of the most variable peaks (Y-axis) across samples (X-axis), with columns arranged by the two clusters identified in (A).

3.5 Identification of Regulatory Rewiring at the EIF5 Locus

To move beyond simple differential expression and identify changes in gene regulation, we performed a differential correlation analysis integrating the transcriptomic (RNA-seq) and chromatin accessibility (ATAC -seq) data.

This analysis identified a significant regulatory rewiring event involving the gene EIF5 (Eukaryotic Translation Initiation Factor 5). In healthy controls, the expression of EIF5 exhibited a significant negative correlation ($r = -0.36$) with a proximal chromatin peak. In the ALS cohort, however, this regulatory relationship was significantly disrupted. The correlation coefficient inverted to a positive value ($r = 0.15$), yielding a highly significant z-difference score of 4.77 ($p = 1.8 \times 10^{-6}$; $q = 0.03$), indicating a statistical inversion of the regulatory logic at this locus.

To infer patient-specific transcription factor (TF) activity, we implemented the PSIONIC framework, which integrates ATAC-seq and RNA-seq data to model gene regulation. To ensure the robustness of the inferred activities, we performed a rigorous hyperparameter optimization, systematically varying the regularization strength and the number of latent regulatory factors.

Model performance was evaluated by computing the Spearman correlation between the predicted gene expression (derived from the inferred TF activities) and the actual measured gene expression from the RNA-seq data. Across all tested parameter configurations, the model's predictive accuracy remained consistently low. The global Spearman correlation between predicted and observed expression failed to exceed 0.14 in any iteration.

Given that this correlation threshold indicates a weak ability to capture the regulatory logic linking chromatin accessibility to gene expression in this dataset, we concluded that the inferred TF activities lacked sufficient fidelity for biological interpretation. Consequently, the analysis was halted at this validation stage, and downstream differential TF activity assessment was not performed.

RESULTS- PART II

3.6 StringDB Analysis of Rewired Transcription Factors reveal Disease Modifiers

To understand the functional protein-protein interactions among the top-ranked transcription factors, the candidates derived from the differential centrality analysis were

provided to STRINGdb. STRINGdb successfully identified a dense cluster of proteins based on our input list in the oligodendrocyte and microglia cell types. To further resolve this network, we applied the Markov Cluster Algorithm (MCL) using the default parameters. The MCL clustering revealed that a remarkably large number of these proteins share highly interconnected interactions, with the vast majority coalescing tightly into a primary, dominant module, which we designated as Cluster 1

Following the identification of this interconnected cluster, we queried the specific list of genes comprising Cluster 1 against eVGeMdb(Singh et al., 2026), a manually curated database of experimentally validated genetic modifiers for neurodegenerative disorders. we found that more than half of the genes within this cluster (particularly those highlighted in the microglia and oligodendrocyte datasets) consisted of biologically validated neuro-disease modifiers (Figure 4.5).

This extraordinarily high concordance is particularly interesting when considering the shared pathophysiological pathways that bridge these diverse neurodegenerative diseases. Multiple neurodegenerative conditions, including ALS, Alzheimer's, and Parkinson's disease, share converging molecular mechanisms such as impaired proteostasis, chronic neuroinflammation, oxidative stress, and mitochondrial dysfunction. The fact that the nodes within Cluster 1 act as shared modifiers across these converging pathways highlights their robust potential as prime, broad-spectrum targets for drug repurposing intervention.

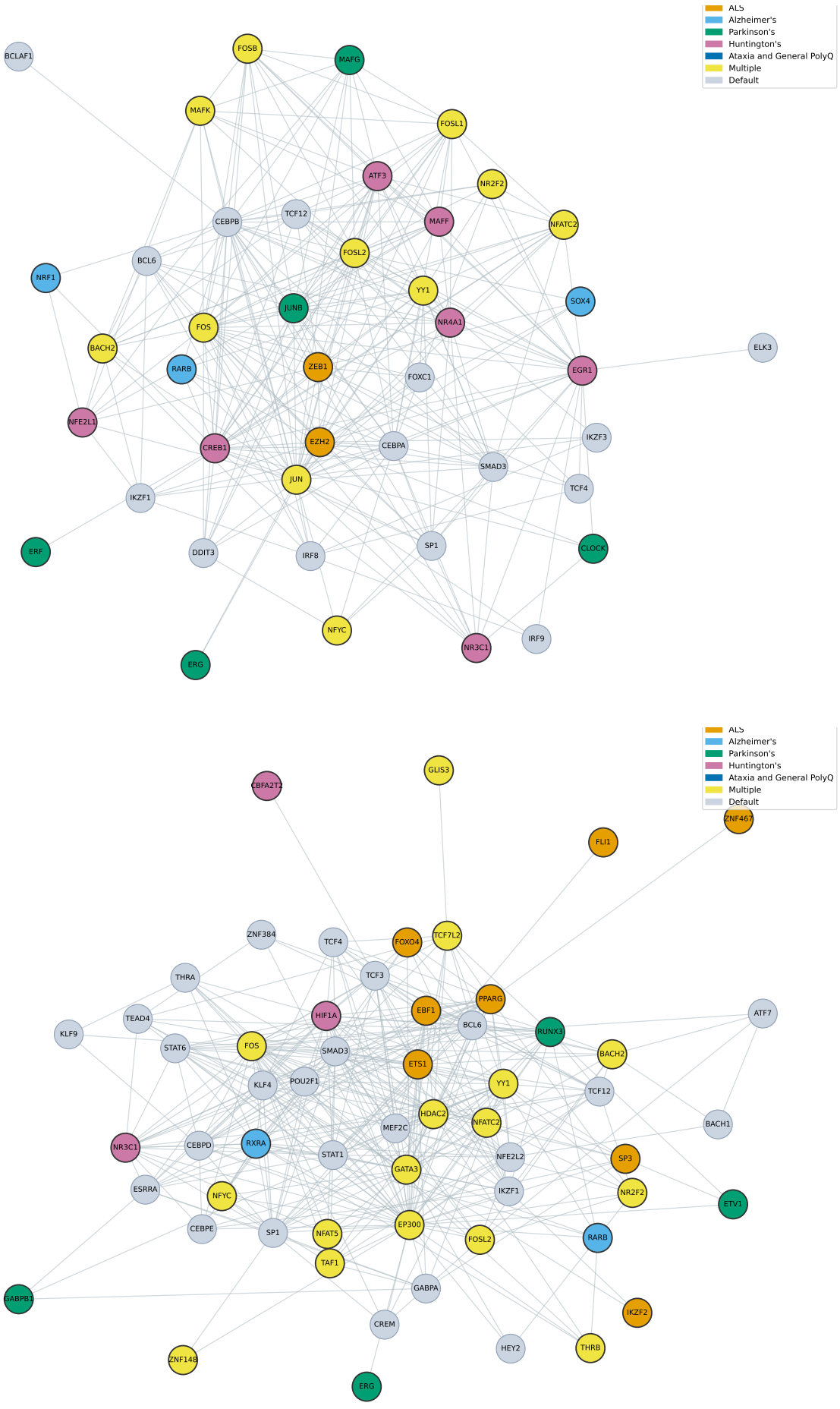
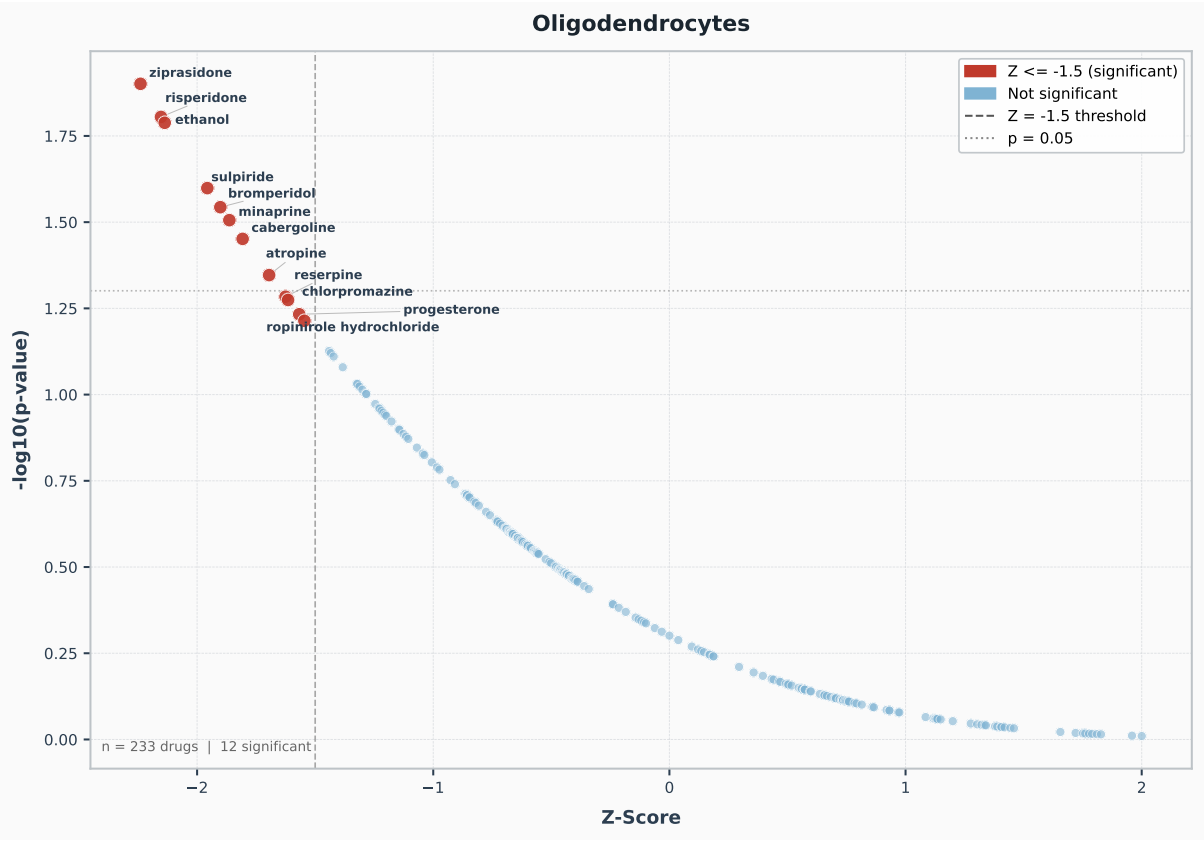


Figure 3.5: Oligodendrocyte and Microglia StringDB Networks of Differentially Rewired Transcription Factors. (A) Oligodendrocyte StringDB network. (B) Microglia StringDB network.

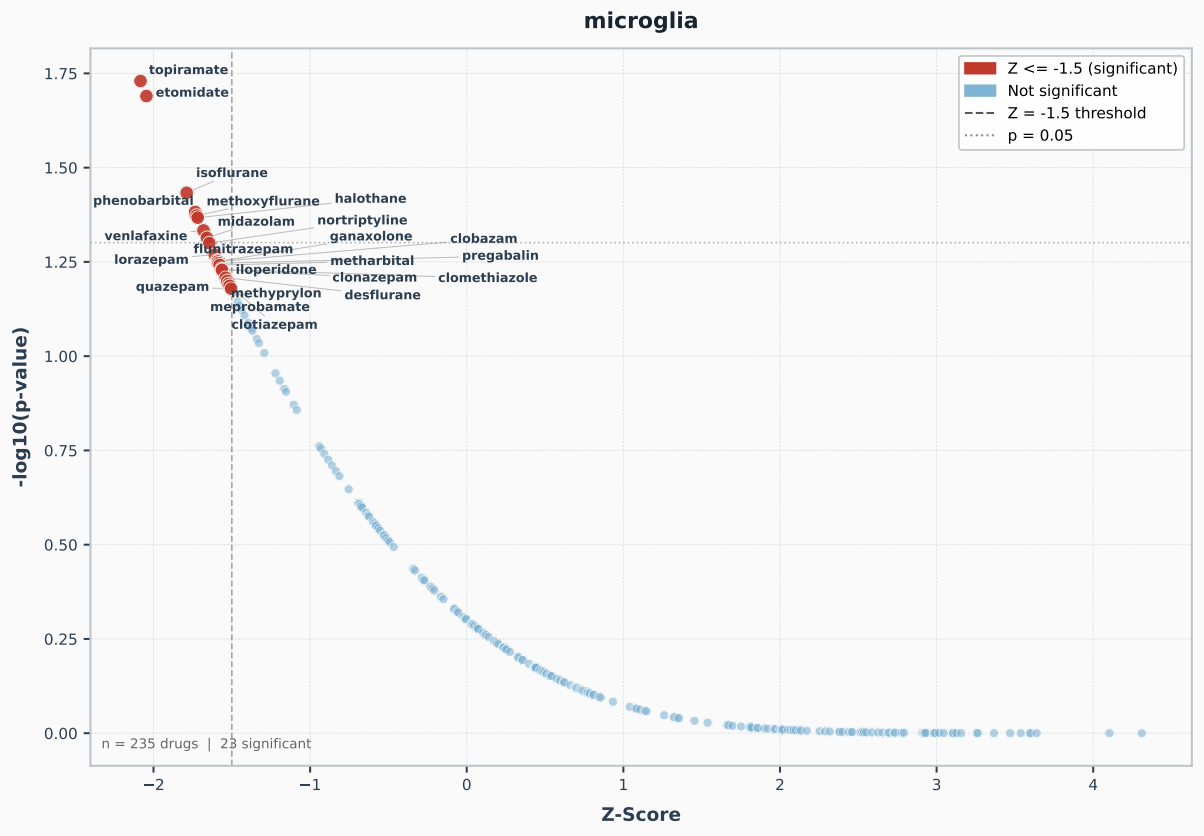
3.7 Drug Repurposing Results

Drug repurposing analysis of the microglial module nominated 23 compounds meeting the significance threshold ($Z \leq -1.5$). The top-ranked compound, topiramate ($Z = -2.082$), targets GRIK2, GRIK4, GRIA2–4, SCN8A/9A, and INSR. Etomidate ($Z = -2.045$) and volatile anaesthetics — isoflurane ($Z = -1.788$), methoxyflurane ($Z = -1.723$), halothane ($Z = -1.717$), and desflurane ($Z = -1.538$) — act as positive allosteric modulators of GABA-A receptor subunits, particularly 2/3-containing receptors. A benzodiazepine cluster comprising midazolam ($Z = -1.659$), lorazepam ($Z = -1.609$), flunitrazepam ($Z = -1.592$), clobazam ($Z = -1.588$), clonazepam ($Z = -1.562$), clotiazepam ($Z = -1.513$), and quazepam ($Z = -1.504$) converged on the α/γ subunit interface of GABA-A receptors. Beyond the GABAergic cluster, venlafaxine ($Z = -1.680$) targeted GRIK4 and SLC6A4/SLC6A2, nortriptyline ($Z = -1.643$) targeted GRIA3 and GRIK2, pregabalin ($Z = -1.583$) targeted CACNA2D1, CACNA1B, and CACNB3, and iloperidone ($Z = -1.566$) targeted NRG3. The astrocyte and excitatory neuron modules exhibited broader topological dispersion, yielding no candidates satisfying the ($Z \leq -1.5$) threshold.

The analysis of the oligodendrocyte module identified several compounds satisfying the predictive threshold ($Z \leq -1.5$), consistent with the observation that the TDP$^{43^{high}}$ cortex exhibits a statistically significant downregulation of myelin-associated genes including MOG, MOBP, and MBP, alongside a high proportion of premature, pre-myelinating oligodendrocytes. The predominant drug class identified was DRD2 antagonists, with risperidone ($Z = -2.124$), pimozide ($Z = -2.063$), sulpiride ($Z = -1.787$), bromperidol ($Z = -1.900$), cabergoline ($Z = -1.807$), chlorpromazine ($Z = -1.614$), and ropinirole hydrochloride ($Z = -1.546$) all satisfying the threshold and converging on DRD2 as a primary gene target. Ziprasidone ($Z = -2.239$), the highest-ranked compound in this module, targets CDH4, a cell adhesion molecule implicated in oligodendrocyte–axon contact. Risperidone additionally targets GRM3, GRID2, and TJP1, implicating glutamate signaling and tight junction integrity within the oligodendrocyte compartment. The steroid hormone progesterone ($Z = -1.906$) also satisfied the threshold, engaging the broadest target profile in this module including NRG1, MAPK1, NFE2L2, TDP1, and the progesterone receptor PGR. Ethanol ($Z = -2.137$) similarly satisfied the threshold, targeting GABA-A and NMDA receptor pathways. The full ranked drug list for this module is presented in Fig. 4.6.



(a) Oligodendrocyte Drug Repurposing results.



(b) Microglia Drug Repurposing results

Figure 3.6: Drug Repurposing Volcano Plots for Oligodendrocytes and Microglia

Chapter 4

Discussion

4.1 Overview of Findings

In the following discussion, we will place the key findings in the context of current ALS research, explore the broader implications of these results for our understanding of ALS pathogenesis, and discuss the potential therapeutic candidates that emerged from our scRNA based network-guided drug repurposing efforts, framing them within the relevant clinical and preclinical evidence available.

4.2 The Confounding Role of Sexual Dimorphism in ALS Transcriptomics

In our initial efforts to map the global transcriptional landscape of ALS using Weighted Gene Co-expression Network Analysis (WGCNA), we faced a significant challenge: the main factors influencing variance were closely linked to the biological sex of participants rather than their disease status. The findings were further supported by Principal Component Analysis (PCA) which showed that the main component (PC1) demonstrated a clear separation between the patients driven entirely by their sex.

Sexual dimorphism in ALS is a well-established observation. Men generally have a higher lifetime risk of developing sporadic ALS compared to women, and there are notable differences in the clinical progression of the disease between the sexes. Recent whole-genome sequencing studies have shed light on this issue, showing that female ALS patients carry a significantly higher load of rare, harmful variants in genes associated with ALS than their male counterparts. This supports the idea of a ”female protective effect,” suggesting that women may require a greater genetic burden to trigger motor neuron degeneration, potentially influenced by unique endocrine factors, like progesterone, or differences in neuro-immune responses (Trojsi et al., 2020; Grassano

et al., 2026).

Our initial WGCNA model, which did not take sex into account, failed to identify any statistically significant disease-related module, highlighting how combining male and female data can obscure important sex-specific signals. However, when we stratified the dataset by sex, we uncovered a unique male-specific co-expression module that was strongly linked to the ALS phenotype. Interestingly, we did not find a similar module for females, suggesting that the transcriptional changes associated with male ALS are highly specific to the male environment. In contrast, female ALS pathology might be more varied or managed by protective mechanisms within the transcriptome.

These insights strongly indicate that biological sex cannot simply be treated as a standard variable in linear models and then regressed out. To accurately identify the different drivers of disease in neurodegenerative research, it's crucial to conduct sex-stratified network analyses.

4.3 Characterization of the Male-Specific Diagnostic Gene Signature

To distill the broad male-specific regulatory network into a clinically viable diagnostic panel, LASSO logistic regression was applied, successfully converging on a sparse, 7-gene signature with a cross-validated AUC of 0.719. Biologically, this module deviates significantly from classical executioner pathways of excitotoxicity or neuroinflammation. Instead, it is overwhelmingly populated by upstream regulatory elements, including non-coding RNAs (ncRNAs), small nucleolar RNAs (snoRNAs), and Zinc Finger (ZNF) proteins. Notably, ncRNAs have emerged as critical regulators of gene expression in ALS, with several long non-coding RNAs and microRNAs previously implicated in motor neuron degeneration, TDP-43 dysfunction, and disease progression (Gagliardi et al., 2012).

A defining feature of the male-specific module is the strong enrichment of non-coding RNAs originating from the DLK1–DIO3 locus on chromosome 14q32, including members of the SNORD113 and SNORD114 snoRNA families (e.g., *SNORD113-6*, *SNORD114-3*) and microRNAs such as *MIR127*, *MIR370*, and *MIR433* (Benetatos et al., 2013). This region is a conserved, maternally imprinted genomic domain that regulates neurodevelopment and cellular differentiation and has been implicated in FUS-associated ALS (Capauto et al., 2018). Given the severe oxidative and metabolic stress experienced by ALS motor neurons, dysregulation of the DLK1–DIO3 cluster may promote abnormal tRNA fragmentation and disrupt translational homeostasis, thereby contributing to proteostatic failure in diseased neurons (van Ingen et al., 2022).

The LASSO signature also identified *FAM50B*, *LINC02301*, and *HSPA2* as positive

predictors of ALS, and *ZNF682* as a negative predictor. Notably, differential expression analyses in *C9orf72* ALS patients have reported significant downregulation of *HSPA2*, and functional studies in *C9orf72* fly models suggest that *HSPA2* exerts a neuroprotective effect that slows neurodegeneration (NeuroLINCS Consortium et al., 2021). FAM50B is an imprinted gene regulated by a promoter-associated differentially methylated region; its increased expression may indicate loss of imprinting or dysregulated methylation states within the ALS epigenetic landscape(Zhang et al., 2011). Long non-coding RNAs can function as molecular scaffolds that recruit chromatin-modifying complexes to specific genomic loci. Dysregulation of such RNA scaffolds can perturb chromatin organization and RNA-binding protein networks, processes that are frequently disrupted in ALS. (Tsai et al., 2010).

4.4 Methodological Artifacts in Subtyping and TF Inference

NMF and Early Stage Tissue Heterogeneity

In our attempt to discover latent transcriptomic subtypes within the cohort, Non-negative Matrix Factorization (NMF) failed to yield stable or biologically relevant clusters. While NMF-based methods have successfully categorized late stage Post-Mortem ALS into distinct transcriptomic subtypes (e.g., ALS-Ox, ALS-Glia, ALS-TE) Tam et al. (2019), the mathematical failure of NMF in our matrix likely stems from the inherent limitations of linear factorization algorithms. Early-stage iPSCs from answerALS tend to display relatively homogeneous transcriptional programs, strong protocol- and batch-driven variability; these features substantially reduce the separability of subtle subtype signatures in bulk expression matrices. Consequently, simple linear factorization cannot reliably disentangle these early, cell-state differences without more expressive non-linear neural-network architectures.

ATAC-seq Mapping Artifacts

A critical methodological observation emerged during the consensus clustering of the ATAC-seq chromatin accessibility data. An initial, highly stable two-cluster structure ($k = 2$) was identified. However, detailed annotation revealed that this separation was driven entirely by accessible regions mapped to the Y chromosome. Because female genomes lack a Y chromosome, these loci naturally show an absence of signal in female samples. Importantly, technical mapping variability in some male samples also resulted in a loss or near absence of signal at Y-linked regions, causing these males to resemble the female accessibility profile. This introduced an artificial clustering structure in which a subset of male samples grouped with female samples based on this technical feature rather than true biological variation . The subsequent collapse of the

clustered structure following removal of Y-chromosome peaks therefore highlights an important methodological consideration for epigenomic analyses: unbiased clustering algorithms must be paired with careful genomic annotation and filtering to ensure that sex-chromosome mapping artifacts are not misinterpreted as biologically meaningful structure

Limitations of PSIONIC

The low predictive performance of PSIONIC observed in this study, reflected by a global Spearman correlation of 0.14 , is most likely attributed to the heterogeneity of the data. PSIONIC works by combining two sources of information: chromatin accessibility data from ATAC-seq and gene expression data from RNA-seq. For the model to infer meaningful TF activity, both of these data types need to show enough variation across samples so that the model can identify consistent patterns linking TF binding to changes in gene expression. When this variation is too small, the model has very little signal to work with, and its ability to make accurate predictions is fundamentally limited. In this cohort, the RNA-seq data is the most likely source of this problem. The samples used in this study were derived from iPSCs under tightly controlled differentiation conditions, meaning that the transcriptional profiles across samples were highly similar to one another. When samples are this similar, the differences in gene expression that PSIONIC relies on to learn TF-target relationships become very small, reducing the range of values the model can use during inference. This low-between sample transcriptional heterogeneity is a well-recognised challenge for regression-based computational frameworks that depend on observing meaningful differences between samples rather than absolute expression levels Osmanbeyoglu et al. (2019b).

4.5 Regulatory Rewiring at the *EIF5* Locus

By integrating RNA-seq with ATAC-seq data, our differential correlation analysis identified a significant inversion of regulatory logic at the *EIF5* locus. In healthy controls, *EIF5* expression exhibited a negative correlation ($r = -0.36$) with a proximal chromatin peak. In ALS, this relationship inverted to a positive correlation ($r = 0.15$, $p = 1.8 \times 10^{-6}$), suggesting that epigenetic repression at this locus had been dismantled.

EIF5 is a GTPase-activating protein for the eIF2·GTP·Met-tRNA$_i^{\mathrm{Met}}$ ternary complex and plays a central role in start-codon selection. Notably, a recent study showed that eIF5 stimulates CUG-initiated poly-GA repeat-associated non-AUG (RAN) translation in cellular and Drosophila models of C9orf72 FTLD/ALS (Gotoh et al., 2024). On that basis, altered EIF5 expression may represent a candidate modifier of pathological RAN translation, although this mechanism has not yet been directly validated in ALS motor

neurons.

4.6 Convergent Glial Modifiers and Topological Intractability

StringDB analysis of the differentially rewired TFs identified a cluster of 77 experimnetal disease modifiers , 35 in microglia and 42 in oligodendrocytes, linked to neurodegenerative diseases, based on data from the eVGeMdb database.(Singh et al., 2026)

Notably, 13 factors including FOS, FOSL1, FOSL2, BACH2, YY1, SERPINE1, NR2F2, NFATC2, and NR3C1 were found in both cell types, indicating that some transcriptional changes in ALS overlap with networks involved in other neurodegenerative diseases. This supports previous findings that glial responses tend to be similar across these conditions, as seen in Disease-Associated Microglia (DAM) and Disease-Associated Oligodendrocytes Raffaele et al. (2021) Jauregui et al. (2023).

Conversely, no repurposed compound met the predictive significance threshold $(Z \leq -1.5)$ within the astrocyte and excitatory neuron networks. This is likely a consequence of the comparatively sparse connectivity of these networks relative to the glial network a less densely rewired topology provides fewer high-confidence protein interaction edges for the drug-target scoring algorithm to act on, reducing the power to detect significant perturbational drug targets. This does not necessarily imply that neuronal and astrocytic pathways are therapeutically irrelevant in ALS, but rather that the current network structure does not provide sufficient resolution to nominate candidates within these compartments.

4.7 Network-Guided Drug Repurposing

Microglia

In ALS, microglia shift from an early neuroprotective state to a chronically activated neurotoxic phenotype that actively contributes to motor neuron death through sustained pro-inflammatory cytokine release and oxidative stress (Liao et al., 2012). Despite being a well-validated pathological mechanism, microglial neuroinflammation has remained a difficult therapeutic target in ALS, underscoring the need for network-informed candidate identification.Drug repurposing analysis of the microglial module nominated 23 compounds meeting the significance threshold (Z 1.5). Notably, many of these drugs target GABA-A receptor subunits, especially GABRA3, GABRA4,

GABRB2, and GABRG3. This finding is significant because microglial cells have functional GABA-A receptors, and activating these receptors has been shown to reduce inflammation in microglia across various studies(Liu et al., 2016) (Bhandage and Barragan, 2021). The convergence of various drug classes such as anesthetics, benzodiazepines, barbiturates, and neurosteroids through our approach on this specific target family supports a coherent mechanism and indicates that modulating GABA receptors to influence microglial inflammation is a computationally valid strategy in ALS.

The top-ranked compound, topiramate (Z = 2.082), targets the broadest node set including GRIK2, GRIK4, GRIA2-4, SCN8A/9A, and INSR, and has been shown to suppress LPS-induced IL-1, IL-6, and TNF- release from primary microglial cultures (Andrzejczak et al., 2016). While a placebo-controlled RCT in 296 ALS patients reported worsening limb strength and excess adverse events at high doses Cudkowicz et al. (2003), Nevertheless, the independent nomination of topiramate through network proximity analysis, consistent with its validated microglial activity, supports the biological relevance of the computational framework, even where clinical translation remains challenging.

Etomidate (Z = 2.045) and the volatile anaesthetics — isoflurane (Z = 1.788), methoxyflurane (Z = 1.723), halothane (Z = 1.717), and desflurane (Z = 1.538) — are positive modulators of GABA receptors and, at clinically relevant concentrations, engage overlapping GABA subunit sets (etomidate being notably selective for 2/3-containing receptors), although exact binding sites and subunit preferences differ between intravenous and inhaled agents. Their relevance to microglial biology is supported by experimental work: isoflurane preconditioning or brief isoflurane exposure attenuates microglial activation and reduces LPS+IFN--induced microglial injury, nitric oxide and glutamate release in murine models/cell systems(Xu et al., 2008). Mechanistically, volatile anaesthetic conditioning has been linked to suppression of TLR4$\rightarrow$NF-B signaling and—through antioxidant/Nrf2-linked pathways—to promotion of anti-inflammatory/M2-like phenotypes in several injury paradigms, though the precise cascade appears context-dependent(Zhang and Yin, 2023). Crucially, general anaesthetics can show both anti- and pro-inflammatory microglial effects depending on dose, duration, age and injury context; therefore the nomination of this class is mechanistically plausible but would require careful dose/exposure calibration and validation in ALS-relevant models(Yang et al., 2024)

The benzodiazepine class — comprising midazolam (Z = 1.659), lorazepam (Z = 1.609), flunitrazepam (Z = 1.592), clobazam (Z = 1.588), clonazepam (Z = 1.562), clotiazepam (Z = 1.513), quazepam (Z = 1.504) largely converge on the central benzodiazepine binding site on GABA receptors (an extracellular pocket at the α/γ subunit interface), although individual agents can show differing subunit selectivity and pharmacology. The collective anti-inflammatory activity of several benzodiazepines

on microglia is supported by direct in vitro evidence: in primary rat microglial cultures clonazepam, midazolam and diazepam markedly decreased LPS-induced nitric oxide release, reduced TNF- secretion, and inhibited microglial proliferation(Wilms et al., 2003)Midazolam also suppresses IL-1-induced IL-6 release from glial cells by inhibiting STAT3 phosphorylation(Tanabe et al., 2011). Because many nominated compounds share nearly identical receptor pharmacology, they often represent a single functional opportunity (a common GABA/benzodiazepine-site or TSPO-mediated mechanism) rather than multiple independent pharmacological leads

The most pharmacologically distinct signals come from outside the GABAergic cluster. Venlafaxine (Z = 1.680) and nortriptyline (Z = 1.643) target glutamatergic nodes GRIK4 and GRIA3/GRIK2 respectively, and both carry well-documented microglial anti-inflammatory activity. SNRIs including venlafaxine and TCAs prevented microglial activation, with studies showing reduced microglial reactivity and decreased immune and oxidative stress products in LPS-stimulated microglial models of neuroinflammation ScienceDirect (Mariani et al., 2022). More specifically, nortriptyline directly inhibits IL-1 and TNF- release from LPS-activated rat mixed glial and microglial cell cultures at concentrations achievable in plasma and brain structures during treatment (Dubovický et al., 2014). Pregabalin (Z = 1.583), targeting calcium channel subunits CACNA2D1, CACNA1B, and CACNB3, provides a mechanistically distinct anti-inflammatory route through suppression of microglial activation via the HMGB1/TLR4/NF-B axis Zhang et al. (2022). Finally, iloperidone (Z = 1.566) targets NRG3, implicating the neuregulin-ErbB signalling axis a pathway with direct relevance to microglial activation and motor neuron survival in ALS models Liu et al. (2018). These four compounds, operating through non-redundant mechanisms, represent the highest-priority candidates for downstream experimental validation within the ALS microglial inflammatory network.

Oligodendrocytes

The mature oligodendrocyte module returned a drug signature dominated by DRD2(based on drug target proximal to the disease module). The dopamine D2 receptor is expressed in a subset of mature interfascicular oligodendrocytes in the rat corpus callosum (Howard et al., 1998), and oligodendrocytes at all maturation stages express mRNA for D3 dopamine receptors (Bongarzone et al., 1998). Loss of D2 receptor signalling has direct consequences for myelin maintenance: D2 receptor knockout mice display a baseline reduction in mature oligodendrocyte numbers in the corpus callosum, and stress-induced myelin loss in wild-type animals is accompanied by a decrease in active -catenin and TCF4-expressing cells among oligodendrocyte lineage cells a regulation entirely absent in D2R-null animals (Choi et al., 2017). The conver-

gence of multiple top-ranked compounds including risperidone, sulpiride, bromperidol, cabergoline, chlorpromazine, and ropinirole hydrochloride on DRD2 as a shared target therefore reflects a pharmacologically and mechanistically coherent signal arising from the disrupted dopaminergic regulation of oligodendrocyte identity in the TDP-43 high ALS cortex.

Among the DRD2-targeting candidates, ropinirole hydrochloride carries the most direct ALS-specific clinical validation. Originally identified as an ALS candidate through iPSC-based motor neuron drug screening, it progressed to the ROPALS phase I/IIa randomised controlled trial. In the open-label extension period, the ropinirole group showed significant suppression of ALSFRS-R decline and an additional 27.9 weeks of disease-progression-free survival, with iPSC-derived motor neurons from participants showing dopamine D2 receptor expression and a potential involvement of the SREBP2-cholesterol pathway in the therapeutic effects (Morimoto et al., 2023).

Cabergoline similarly exerts neuroprotection through D2 receptor-mediated mechanisms: cabergoline prevented oxidative stress-induced death of cultured cortical neurons via a D2 receptor-dependent mechanism, suppressing ERK1/2 activation and reducing extracellular glutamate accumulation, an effect abolished in the presence of a D2 receptor antagonist (Odaka et al., 2014). Given that oxidative stress and glutamate excitotoxicity are cardinal features of ALS pathology, the identification of these compounds within the mature oligodendrocyte module points to a convergent vulnerability of dopaminergic signalling in this cell population.

Risperidone and ziprasidone, while lacking direct ALS oligodendrocyte studies, have been independently nominated through structure-based computational approaches as the top ALS candidate compounds capable of inhibiting several ALS-associated proteins including casein kinase 1 (CK1), protein tyrosine kinase 2 (PTK2/FAK), and ephrin type-A receptor 4 (EPHA4).(Dogra et al., 2025). Risperidone's additional targets — GRM3, GRID2, and TJP1 — implicate broader dysregulation of glutamate signalling and tight junction integrity within the oligodendrocyte compartment, while ziprasidone's top target CDH4 points to impaired axo-glial communication, consistent with the axonal pathology observed in ALS.

Progesterone presents the most direct myelin-relevant mechanistic evidence among all candidates in this module. In preclinical spinal cord injury models, progesterone increased oligodendrocyte precursor cell numbers, upregulated myelin basic protein expression, and enhanced expression of Olig2 and Nkx2.2 — transcription factors central to oligodendrocyte specification and differentiation (Labombarda et al., 2009). Prolonged progesterone treatment further enhanced Olig1 expression, a transcription factor specifically required for myelin lesion repair, while simultaneously restoring molecular markers of motoneuron function (Labombarda et al., 2011). Its broad target profile — engaging NRG1, MAPK1, NFE2L2 further positions progesterone as the

most biologically coherent therapeutic candidate for restoring mature oligodendrocyte function in the TDP-43 high ALS cortex

4.8 Conclusion

The findings presented in this thesis provide important biological insights while also revealing significant weaknesses in the methods used in computational neurogenomics. Our research shows that unsupervised global network analyses are especially sensitive to major biological factors, with sexual dimorphism being a key example; if not properly accounted for, these factors can completely mask disease-specific signals. By carefully stratifying cohorts and correcting for artifacts, we successfully identified a male-associated diagnostic gene signature that is primarily influenced by upstream non-coding RNAs and epigenetic regulators at the 14q32 locus, rather than by downstream inflammatory factors.

Additionally, our integration of multi-omic data revealed a critical regulatory inversion at the EIF5 locus. Furthermore, using cell-type-specific network proximity algorithms, we highlighted the different vulnerabilities found in glial cells and neurons during the later stages of ALS. This underscores the importance of targeting microglial and oligodendrocyte networks for potential therapeutic interventions involving repurposed immunomodulatory and remyelinating treatments.

Together, these results suggest that ALS should not be viewed as a disease defined by a single molecular issue, but rather as one marked by overlapping regulatory problems across transcriptional, epigenomic, and network levels. This dysfunction is specific to certain cell types, varies by sex, and changes depending on the disease stage. Our work highlights that rigorous analysis is a vital biological tool: the most relevant disease signals are often hidden not because they are absent, but due to limitations in the frameworks we use to detect them. Moving forward, developing ALS therapeutics will require computational approaches that reflect the complexity of the disease itself.

Bibliography

Andrzejczak, D., Woldan-Tambor, A., Bednarska, K., and Zawilska, J. B. (2016). The effects of topiramate on lipopolysaccharide (lps)-induced proinflammatory cytokine release from primary rat microglial cell cultures. *Epilepsy Research*, 127:352–357.

Baxi, E. G., Thompson, T., Li, J., Kaye, J. A., Lim, R. G., Wu, J., Ramamoorthy, D., Lima, L., Vaibhav, V., Matlock, A., Frank, A., Coyne, A. N., Landin, B., Ornelas, L., Mosmiller, E., Thrower, S., Farr, S. M., Panther, L., Gomez, E., Galvez, E., Perez, D., Meepe, I., Lei, S., Mandefro, B., Trost, H., Pinedo, L., Banuelos, M. G., Liu, C., Moran, R., Garcia, V., Workman, M., Ho, R., Wyman, S., Roggenbuck, J., Harms, M. B., Stocksdale, J., Miramontes, R., Wang, K., Venkatraman, V., Holewenski, R., Sundararaman, N., Pandey, R., Manalo, D. M., Donde, A., Huynh, N., Adam, M., Wassie, B. T., Vertudes, E., Amirani, N., Raja, K., Thomas, R., Hayes, L., Lenail, A., Cerezo, A., Luppino, S., Farrar, A., Pothier, L., Prina, C., Morgan, T., Jamil, A., Heintzman, S., Jockel-Balsarotti, J., Karanja, E., Markway, J., McCallum, M., Joslin, B., Alibazoglu, D., Kolb, S., Ajroud-Driss, S., Baloh, R., Heitzman, D., Miller, T., Glass, J. D., Patel-Murray, N. L., Yu, H., Sinani, E., Vigneswaran, P., Sherman, A. V., Ahmad, O., Roy, P., Beavers, J. C., Zeiler, S., Krakauer, J. W., Agurto, C., Cecchi, G., Bellard, M., Raghav, Y., Sachs, K., Ehrenberger, T., Bruce, E., Cudkowicz, M. E., Maragakis, N., Norel, R., Van Eyk, J. E., Finkbeiner, S., Berry, J., Sareen, D., Thompson, L. M., Fraenkel, E., Svendsen, C. N., and Rothstein, J. D. (2022). Answer als, a large-scale resource for sporadic and familial als combining clinical and multi-omics data from induced pluripotent cell lines. *Nature Neuroscience*, 25(2):226–237.

Benetatos, L., Hatzimichael, E., Londin, E., Vartholomatos, G., Loher, P., Rigoutsos, I., and Briasoulis, E. (2013). The micrornas within the dlk1-dio3 genomic region: involvement in disease pathogenesis. *Cellular and Molecular Life Sciences*, 70(5):795–814. Epub 2012 Jul 24.

Bhandage, A. K. and Barragan, A. (2021). Gabaergic signaling by cells of the immune system: more the rule than the exception. *Cellular and Molecular Life Sciences*, 78(15):5667–5679.

Bongarzone, E. R., Howard, S. G., Schonmann, V., and Campagnoni, A. T. (1998).

Identification of the dopamine d3 receptor in oligodendrocyte precursors: potential role in regulating differentiation and myelin formation. *Journal of Neuroscience*, 18(14):5344–5353.

Capauto, D., Colantoni, A., Lu, L., Santini, T., Peruzzi, G., Biscarini, S., Morlando, M., Shneider, N. A., Caffarelli, E., Laneve, P., and Bozzoni, I. (2018). A regulatory circuitry between *gria2*, mir-409, and mir-495 is affected by als *fus* mutation in esc-derived motor neurons. *Molecular Neurobiology*, 55(10):7635–7651. Epub 2018 Feb 12.

Choi, M.-H., Na, J. E., Yoon, Y. R., Lee, H. J., Yoon, S., Rhyu, I. J., and Baik, J.-H. (2017). Role of dopamine d2 receptor in stress-induced myelin loss. *Scientific Reports*, 7:11654.

Choudhury, C., Murugan, N. A., and Priyakumar, U. D. (2022). Structure-based drug repurposing: Traditional and advanced ai/ml-aided methods. *Drug Discovery Today*, 27(7):1847–1861.

Cudkowicz, M. E., Shefner, J. M., Schoenfeld, D. A., Brown, R. H. J., Johnson, H., Qureshi, M., Jacobs, M., Rothstein, J. D., Appel, S. H., Pascuzzi, R. M., Heiman-Patterson, T. D., Donofrio, P. D., David, W. S., Russell, J. A., Tandan, R., Pioro, E. P., Felice, K. J., Rosenfeld, J., Mandler, R. N., Sachs, G. M., Bradley, W. G., Raynor, E. M., Baquis, G. D., Belsh, J. M., Novella, S., Goldstein, J., Hulihan, J., and Consortium, N. A. (2003). A randomized, placebo-controlled trial of topiramate in amyotrophic lateral sclerosis. *Neurology*, 61(4):456–464.

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of r&d costs. *Journal of Health Economics*, 47:20–33.

Dogra, S., Kouznetsova, V. L., and Tsigelny, I. F. (2025). Repurposing fda-approved drugs for treatment of amyotrophic lateral sclerosis using machine learning. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 27(1-2):1–9.

Dubovický, M., Császár, E., Melicherčíková, K., Kuniakóvá, M., and Račková, L. (2014). Modulation of microglial function by the antidepressant drug venlafaxine. *Interdisciplinary Toxicology*, 7(4):201–207.

Franzén, O., Gan, L.-M., and Björkegren, J. L. M. (2019). Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database*, 2019:baz046.

Freshour, S. L., Kiwala, S., Cotto, K. C., Coffman, A. C., McMichael, J. F., Song, J. J., Griffith, M., and Wagner, A. H. (2021). Integration of the drug–gene interaction database (dgidb 4.0) with open crowdsource efforts. *Nucleic Acids Research*, 49(D1):D1144–D1151.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Gagliardi, S., Milani, P., Sardone, V., Pansarasa, O., and Cereda, C. (2012). From transcriptome to noncoding rnas: Implications in als mechanism. *Neurology Research International*, 2012:278725.

Gaujoux, R. and Seoighe, C. (2010). A flexible r package for nonnegative matrix factorization. *BMC Bioinformatics*, 11:367.

Geloso, M. C., Corvino, V., Marchese, E., Serrano, A., Michetti, F., and D'Ambrosi, N. (2017). The dual role of microglia in als: mechanisms and therapeutic approaches. *Cellular and Molecular Life Sciences*, 74(16):2877–2895.

Gotoh, S., Mori, K., Fujino, Y., Kawabe, Y., Yamashita, T., Omi, T., Nagata, K., Tagami, S., Nagai, Y., and Ikeda, M. (2024). eif5 stimulates the cug initiation of ran translation of poly-ga dipeptide repeat protein (dpr) in c9orf72 ftld/als. *J. Biol. Chem.*, 300(3):105703.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018.

Grassano, M., Palumbo, F., Mora, G., Gallone, S., De Marco, G., Merulla, I., Paolantonio, C., Maccabeo, A., Canosa, A., Manera, U., Vasta, R., Iazzolino, B., Testa, M., Fuda, G., Salamone, P., Marchese, G., Casale, F., Moglia, C., Calvo, A., Borghero, G., and Chiò, A. (2026). Sex-specific genetic architecture of als: Evidence of a female protective effect? *Annals of Neurology*. Online ahead of print.

Gupta, C., Cohen Kalafut, N., Clarke, D., Choi, J. J., Hanthanan Arachchilage, K., Khullar, S., Xia, Y., Zhou, X., Dursun, C., Gerstein, M., and Wang, D. (2025). Network-based drug repurposing for psychiatric disorders using single-cell genomics. *Cell Genomics*, 5(11):101003.

Howard, S. G., Landry, C., Fisher, R., Bezouglaia, O., Handley, V., and Campagnoni, A. T. (1998). Postnatal localization and morphogenesis of cells expressing the dopaminergic d2 receptor gene in rat brain: expression in non-neuronal cells. *Journal of Comparative Neurology*, 391(1):87–98.

Jauregui, C., Blanco-Luquin, I., Macías, M., Roldan, M., Caballero, C., Pagola, I., Mendioroz, M., and Jericó, I. (2023). Exploring the disease-associated microglia state in amyotrophic lateral sclerosis. *Biomedicines*, 11(11):2994.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296.

Kumar, A. and Daumé, H. I. (2012). Learning task grouping and overlap in multi-task learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1383–1390, Edinburgh, Scotland. Omnipress. GO-MTL.

Labombarda, F., Gonzalez, S., Lima, A., Roig, P., Guennoun, R., Schumacher, M., and De Nicola, A. F. (2011). Progesterone attenuates astro- and microgliosis and enhances oligodendrocyte differentiation following spinal cord injury. *Experimental Neurology*, 231:135–146.

Labombarda, F., Gonzalez, S. L., Lima, A., Roig, P., Guennoun, R., Schumacher, M., and de Nicola, A. F. (2009). Effects of progesterone on oligodendrocyte progenitors, oligodendrocyte transcription factors, and myelin proteins following spinal cord injury. *Glia*, 57(8):884–897.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., and Golub, T. R. (2006). The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559.

Liao, B., Zhao, W., Beers, D. R., Henkel, J. S., and Appel, S. H. (2012). Transformation from a neuroprotective to a neurotoxic microglial phenotype in a mouse model of ALS. *Experimental Neurology*, 237(1):147–152.

Liu, H., Leak, R. K., and Hu, X. (2016). Neurotransmitter receptors on microglia. *Stroke and Vascular Neurology*, 1(2):52–58.

Liu, J., Bhatt, D. L., and Song, F. (2018). Slowing disease progression in the SOD1 mouse model of ALS by blocking neuregulin-induced microglial activation. *Neurobiology of Disease*, 111:1–11.

Mariani, N., Everson, J., Pariante, C. M., and Borsini, A. (2022). Modulation of microglial activation by antidepressants. *Journal of Psychopharmacology*, 36(2):131–150.

Masrori, P. and Van Damme, P. (2020). Amyotrophic lateral sclerosis: a clinical review. *European Journal of Neurology*, 27(10):1918–1929.

Meng, F., Xi, Y., Huang, J., and Ayers, P. W. (2021). A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors. *Scientific Data*, 8:289.

Morimoto, S., Takahashi, S., Ito, D., Date, Y., Okano, H., and et al. (2023). Phase 1/2a clinical trial in als with ropinirole, a drug candidate identified by ipsc drug discovery. *Cell Stem Cell*, 30(6):766–780.e9.

Muzellec, B., Teleńczuk, M., Cabeli, V., and Andreux, M. (2023). Pydeseq2: a python package for bulk rna-seq differential expression analysis. *Bioinformatics*, 39(9):btad547.

NeuroLINCS Consortium, NYGC ALS Consortium, Li, J., Lim, R. G., Kaye, J. A., Dardov, V., Coyne, A. N., Wu, J., Milani, P., Cheng, A., Thompson, T. G., Ornelas, L., Frank, A., Adam, M., Banuelos, M. G., Casale, M., Cox, V., Escalante-Chong, R., Daigle, J. G., Gomez, E., Hayes, L., Holewenski, R., Lei, S., Lenail, A., Lima, L., Mandefro, B., Matlock, A., Panther, L., Patel-Murray, N. L., Pham, J., Ramamoorthy, D., Sachs, K., Shelley, B., Stocksdale, J., Trost, H., Wilhelm, M., Venkatraman, V., Wassie, B. T., Wyman, S., Yang, S., and Svendsen, C. N. (2021). An integrated multiomic analysis of ipsc-derived motor neurons from c9orf72 als patients. *iScience*, 24(11):103221.

Odaka, H., Numakawa, T., Adachi, N., Ooshima, Y., Nakajima, S., Katanuma, Y., Inoue, T., and Kunugi, H. (2014). Cabergoline, dopamine d2 receptor agonist, prevents neuronal cell death under oxidative stress via reducing excitotoxicity. *PLOS ONE*, 9(6):e99271.

Osmanbeyoglu, H. U., Shimizu, F., Rynne-Vidal, A., Alonso-Curbelo, D., Chen, H.-A., Wen, H. Y., Yeung, T.-L., Jelinic, P., Razavi, P., Lowe, S. W., Mok, S. C., Chiosis, G., Levine, D. A., and Leslie, C. S. (2019a). Chromatin-informed inference of transcriptional programs in gynecologic and basal breast cancers. *Nature Communications*, 10:4369. Open Access.

Osmanbeyoglu, H. U., Shimizu, F., Rynne-Vidal, A., Ayala, D., Barthorpe, A., et al. (2019b). Chromatin-informed inference of transcriptional programs in gynecologic and basal breast cancers. *Nature Communications*, 10:4369.

Pascual-Montano, A., Carazo, J. M., Kochi, K., Lehmann, D., and Pascual-Marqui, R. D. (2006). Nonsmooth nonnegative matrix factorization (nsnmf). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):403–415.

Philips, T. and Rothstein, J. D. (2013). Oligodendrocyte dysfunction in the pathogenesis of amyotrophic lateral sclerosis. *Nature Neuroscience*, 17(1):18–19.

Raffaele, S., Boccazzi, M., and Fumagalli, M. (2021). Oligodendrocyte dysfunction in amyotrophic lateral sclerosis: Mechanisms and therapeutic perspectives. *Cells*, 10(3):565.

Singh, S., Singh, P., Varada, M. P., Vijaivasudev, M. J., Kumar, R., and Malik, I. (2026). evgemdb: a manually curated database for experimentally validated genetic modifiers of neurodegenerative disorders. *NAR Molecular Medicine*, 3(1):ugaf044.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., Lahr, D. L., Hirschman, J. E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I. C., Lam, D., Liberzon, A., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17.

Tam, O. H., Rozhkov, N. V., Shaw, R., Kim, D., Hubbard, I., Fennessey, S., Propp, N., Fagegaltier, D., Harris, B. T., Ostrow, L. W., Phatnani, H., Ravits, J., Dubnau, J., Hammell, M. G., the NYGC ALS Consortium, and et al. (2019). Postmortem cortex samples identify distinct molecular subtypes of als: Retrotransposon activation, oxidative stress, and activated glia. *Cell Reports*, 29(5):1164–1177.e5.

Tanabe, K., Kozawa, O., and Iida, H. (2011). Midazolam suppresses interleukin-1ß-induced interleukin-6 release from rat glial cells. *Journal of Neuroinflammation*, 8:68.

Trojsi, F., D'Alvano, G., Bonavita, S., and Tedeschi, G. (2020). Genetics and sex in the pathogenesis of amyotrophic lateral sclerosis (als): Is there a link? *International Journal of Molecular Sciences*, 21(10):3647.

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., Shi, Y., Segal, E., and Chang, H. Y. (2010). Long noncoding rna as modular scaffold of histone modification complexes. *Science*, 329(5992):689–693.

Tsitkov, S., Valentine, K., Kozareva, V., Donde, A., Frank, A., Lei, S., the Answer ALS Consortium, Van Eyk, J. E., Finkbeiner, S., Rothstein, J. D., Thompson, L. M., Sareen, D., Svendsen, C. N., Fraenkel, E., et al. (2024). Disease related changes in atac-seq of ipsc-derived motor neuron lines from als patients and controls. *Nature Communications*, 15:3606. Published 02 May 2024; Open access.

van Ingen, E., Engbers, P. A. M., Woudenberg, T., van der Bent, M. L., Mei, H., Wojta, J., Quax, P. H. A., and Nossent, A. Y. (2022). C/d box snorna snord113-6 guides 2'-o-methylation and protects against site-specific fragmentation of trna^{Leu}(taa) in vascular remodeling. *Molecular Therapy: Nucleic Acids*, 30:162–172.

Wang, H. V., Xiang, J. F., Yuan, C., Veire, A. M., Gendron, T. F., Murray, M. E., Tansey, M. G., Hu, J., Gearing, M., Glass, J. D., Jin, P., Corces, V. G., and McEachin, Z. T. (2025). pTDP-43 levels correlate with cell type specific molecular alterations in the prefrontal cortex of C9orf72 ALS/FTD patients. *Proceedings of the National Academy of Sciences of the United States of America*, 122(9):e2419818122.

Wilkerson, M. D. and Hayes, D. N. (2010). Consensusclusterplus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12):1572–1573.

Wilms, H., Claasen, J., Röhl, C., Sievers, J., Deuschl, G., and Lucius, R. (2003). Involvement of benzodiazepine receptors in neuroinflammatory and neurodegenerative diseases: evidence from activated microglial cells in vitro. *Neurobiology of Disease*, 14(3):417–424.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2006). Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(Database issue):D668–D672.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15.

Workman, M. J., Lim, R. G., Wu, J., Frank, A., Ornelas, L., Panther, L., Galvez, E., Perez, D., Meepe, I., Lei, S., Valencia, V., Gomez, E., Liu, C., Moran, R., Pinedo, L., Tsitkov, S., Ho, R., Kaye, J. A., Thompson, T., Rothstein, J. D., Finkbeiner, S., Fraenkel, E., Sareen, D., Thompson, L. M., and Svendsen, C. N. (2023). Large-scale differentiation of ipsc-derived motor neurons from als and control subjects. *Neuron*, 111(8):1191–1204.e5.

Xu, X., Kim, J. A., and Zuo, Z. (2008). Isoflurane preconditioning reduces mouse microglial activation and injury induced by lipopolysaccharide and interferon-gamma. *Neuroscience*, 154(3):1002–1008.

Yamanaka, K., Chun, S. J., Boillee, S., Fujimori-Tonou, N., Yamashita, H., Gutmann, D. H., Takahashi, R., Misawa, H., and Cleveland, D. W. (2008). Astrocytes as determinants of disease progression in inherited amyotrophic lateral sclerosis. *Nature Neuroscience*, 11(3):251–253.

Yang, Y. et al. (2024). Effect of general anesthetic agents on microglia. *Aging and Disease*, 15(3):1308–1328.

Zhang, A., Skaar, D. A., Li, Y., Huang, D., Price, T. M., Murphy, S. K., and Jirtle, R. L. (2011). Novel retrotransposed imprinted locus identified at human 6p25. *Nucleic Acids Research*, 39(13):5388–5400. Epub 2011 Mar 18.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 17.

Zhang, M. and Yin, Y. (2023). Dual roles of anesthetics in postoperative cognitive dysfunction: Regulation of microglial activation through inflammatory signaling pathways. *Frontiers in Immunology*, 14:1102312.

Zhang, Z., Jiang, J., He, Y., Cai, J., Xie, J., Wu, M., Xing, M., Zhang, Z., Chang, H., Yu, P., Chen, S., Yang, Y., Shi, Z., Liu, Q., Sun, H., He, B., Zeng, J., Huang, J., Chen, J., Li, H., Li, Y., Lin, W.-J., and Tang, Y. (2022). Pregabalin mitigates microglial activation and neuronal injury by inhibiting HMGB1 signaling pathway in radiation-induced brain injury. *Journal of Neuroinflammation*, 19(1):231.

Zhu, L. J., Gazin, C., Lawson, N. D., Pagès, H., Lin, S. M., Lapointe, D. S., and Green, M. R. (2010). Chippeakanno: a bioconductor package to annotate chip-seq and chip-chip data. *BMC Bioinformatics*, 11:237.