



**Assessing the Added Value of scATAC-seq Data
in Cell-Specific Gene Regulatory Network
Inference for Alzheimer's Disease**

Leo Lin¹

Supervisors: Marcel Reinders¹, Bram Pronk¹, Timo Verlaan¹

¹EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Leo Lin

Final project course: CSE3000 Research Project

Thesis committee: Marcel Reinders, Bram Pronk, Timo Verlaan, Sicco Verwer

Abstract

Alzheimer’s disease (AD) is a complex neurodegenerative disorder and the most common form of dementia, accounting for an estimated 60–70% of all cases worldwide [1]. Understanding how gene regulation changes in specific brain cell types during disease progression is essential to uncovering the molecular mechanisms of AD. Recent research has introduced new methods to extend Gene Regulatory Network (GRN) inference with ATAC-seq in addition to RNA-seq, including ScReNI [2].

This paper investigates whether integrating chromatin accessibility into GRN inference can help explain gene regulatory changes in specific brain cell types during AD progression by evaluating whether ScReNI’s ATAC-derived terms improve biological support beyond RNA-only inference. Using a Python reimplement, we decompose the published ScReNI formula and assess formula variants on the mouse retina development benchmark with ChIP-Atlas precision and network-clustering ARI. We then apply the same component analysis to SEA-AD MTG data and evaluate whether literature reported AD-related microglia genes survive feature selection.

The results suggest that removing the regulator-locus term and using TF-specific target-peak attribution improves performance of the ScReNI weight formula on the mouse retina development dataset. However, before ScReNI can support strong AD-specific regulatory claims in SEA-AD, feature selection must retain disease-relevant regulators and validation must be performed using human brain cell-type-specific benchmarks.

1 Introduction

Dementia is a syndrome associated with a decline in cognitive abilities. The most common form of dementia is the complex neurodegenerative disorder Alzheimer’s disease (AD), characterized by the accumulation of amyloid plaque and neurofibrillary tangles [3]. Although amyloid-targeting antibodies have shown a significant slowing of cognitive decline in early symptomatic AD, the molecular events that drive neurodegeneration in specific brain cell types remain unclear [4; 5].

Network-level AD studies have shown that gene modules and molecular subtypes can point to disease mechanisms that are not visible from individual differentially expressed genes alone [6; 7]. Cell-type-specific regulatory maps are also important because AD risk and pathology involve distinct neuronal and glial responses [8; 9]. Single-nucleus studies identify cell-type-dependent disease responses, while chromatin and enhancer-promoter maps connect disease-associated variants to specific brain cell types [10; 11]. Gene regulatory networks (GRNs) provide a framework for studying these cell-type-specific regulatory mechanisms by representing inferred regulatory relationships between transcription factors and target genes.

A GRN is a directed graph in which nodes represent genes and edges encode inferred regulatory interactions. Traditionally, GRNs were built from bulk RNA sequencing data, which averages expression across many cells and masks cell-type-level

variation [12]. However, even within a single annotated cell type, regulatory programs can vary across disease state and local cellular context. Single-cell RNA sequencing (scRNA-seq) addresses this limitation by measuring gene expression in individual cells, allowing for the inference of cell-specific GRNs that can reveal regulatory differences between cell types and disease states [8; 13].

While scRNA-seq measures transcript abundance, Single-cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) measures chromatin accessibility. scATAC-seq identifies genomic regions where chromatin is in an open conformation, allowing transcription factors (TFs) to access and potentially bind to their target sequence motifs [14]. Figure 1 illustrates this relationship: ATAC-seq detects accessible euchromatin as peaks near active regulatory elements and expressed genes, whereas compact heterochromatin remains inaccessible and transcriptionally silent.

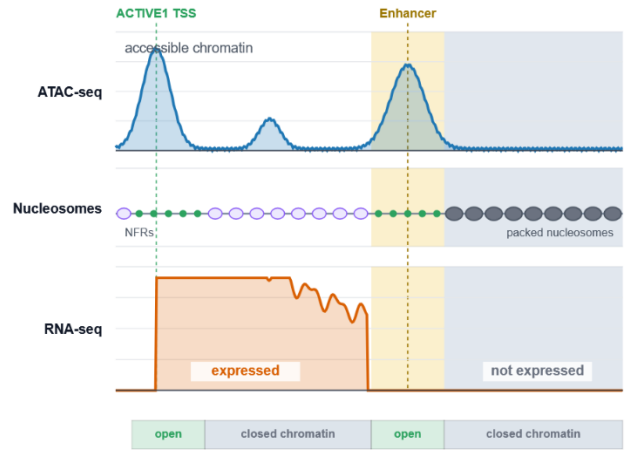


Figure 1: ATAC-seq peaks occur at open nucleosome-depleted regions near active transcription start sites and an enhancer, while the compact heterochromatin region lacks accessibility and shows no ATAC-seq signal.

Single-cell regulatory network inference (ScReNI) is a recently published method for cell-specific GRN inference that extends this approach by integrating scRNA-seq and scATAC-seq [2]. ScReNI trains a random forest model per gene, using the expression of all other genes and the accessibility of nearby chromatin peaks as predictors, then converts the resulting feature importance scores into a weighted, directed GRN for each cell. ScReNI was chosen for this research because it is designed for both paired and unpaired scRNA-seq/scATAC-seq data. By contrast, enhancer-aware alternatives such as SCENIC+ and Pando are primarily designed around paired or cell-state-matched multiome data [15; 16]. This distinction matters because paired measurements require specialized multiome protocols, such as sci-CAR and SHARE-seq [17; 18]. Since many existing disease datasets contain unmatched single-modality measurements, methods capable of integrating both paired and unpaired cells enable GRN inference on datasets that lack paired measurements [19].

Whether adding the ATAC signal actually changes the resulting networks in a meaningful way is the central question of this paper. First, we quantify the contribution of each ATAC component to the final edge weight in wScReNI and how this varies across cell types. Next, we compare the top-

ranked edge sets from the network inferred using only scRNA-seq data (kScReNI) and after incorporating scATAC-seq data (wScReNI) to determine whether adding ATAC features preserves or replaces the RNA-only network topology. We then evaluate whether the resulting top-ranked mouse retina development edges have more ChIP-Atlas TF-binding support. Because chromatin accessibility and enhancer-promoter architecture are cell-type specific in the brain [10; 11], ATAC-derived features may not contribute uniformly across cell types. Separating regulator-locus accessibility from motif-gated target-peak attribution can therefore show which type of chromatin signal is most informative for GRN inference.

The main conclusion of this paper is that a target-specific use of ATAC-derived peak importance improves ChIP-Atlas precision at top-500 in the mouse retina development benchmark among the tested variants, but the published wScReNI weight formula does not behave as a small chromatin correction to the RNA-only kScReNI network. In both MRD and SEA-AD, the full wScReNI formula produces a top-ranked network that is far from kScReNI. The mouse retina development validation suggests that the regulator-locus peak term ($r_{i,j}$) contributes to this topology shift without improving ChIP-Atlas precision in this benchmark, while the target-specific formula ($s_{i,j}$) keeps the ATAC-aware random-forest fit and gives the highest ChIP-Atlas precision among the tested variants.

The rest of the paper is organized as follows. Section 2 describes the datasets, ScReNI pipeline and evaluation metrics. Section 3 decomposes the formula components, evaluates network topology and ChIP-Atlas precision. Sections 4-5 discuss the implications and what must change before the method can support stronger AD regulatory claims.

2 Background and Methodology

2.1 Datasets

The mouse retina development (MRD) dataset is the benchmark used by ScReNI and serves as the main validation dataset in this work [2; 20]. The benchmark subset contains 100 cells for Müller glia (MG) and each of the three retinal progenitor cell stages (RPC1, RPC2 and RPC3). In the feature selection step, all genes are ranked by variability using Seurat v3 and only the top-500 highly variable genes (HVGs) are used by ScReNI to build the GRNs [21]. In the decomposition experiments, 40 cells are sampled from this processed dataset to reduce runtime, with 10 cells per retinal cell type. The gene-peak-TF preprocessing for this input retains 44 TF genes, with 217 unique peaks and 9,954 rows in the triplet table.

Dataset	Cells used	HVGs	TFs	Triplets
MRD	40 of 400	500	44	9,954
SEA-AD MTG	72 of 1,200	500	26	10,424
SEA-AD Microglia	3,129	2,000	-	-

Table 1: Dataset inputs used in the final analyses. The first two rows are the 500-HVG GRN inputs used for formula decomposition and topology evaluation. The microglia row is the separate HVG scope experiment used to test whether AD literature genes survive feature selection.

The Seattle Alzheimer’s Disease Brain Cell Atlas (SEA-AD) provides human middle temporal gyrus (MTG) single-nucleus RNA and ATAC data across Alzheimer’s disease progression [22]. Although ScReNI can also be applied to unpaired

RNA and ATAC inputs, this work uses only paired multiome cells so that RNA and ATAC measurements come from the same nuclei. For ScReNI preprocessing, we sampled 1,200 cells across 24 cortical subclasses, 500 HVGs, 26 TF genes and 10,424 gene-peak-TF triplet rows. For the decomposition experiments reported in the results, 72 cells are sampled from this processed dataset, with three cells per subclass. SEA-AD is used to test whether the component-level patterns observed in mouse retina also appear in human cortex, especially the dominance of the regulator-locus peak term and the topology shift between kScReNI and wScReNI.

In addition to the 500-HVG SEA-AD network input, a separate microglia-specific feature-selection experiment is used to evaluate whether known AD-relevant microglia literature relations can enter the ScReNI pipeline. This experiment filters the paired SEA-AD data to all 3,129 cells annotated as Microglia-PVM, reads raw RNA counts from the UMIs layer and selects 2,000 microglia-specific HVGs from 36,601 measured genes using Seurat v3 VST.

2.2 ScReNI Pipeline Overview

ScReNI infers cell-specific gene regulatory networks by integrating scRNA-seq and scATAC-seq data [2]. Figure 9 in Appendix A shows the original ScReNI workflow schematic. Although the ScReNI paper describes the main inference workflow after preprocessing, we also make feature selection explicit here because it determines which genes and peaks can enter every later step. The wScReNI variant combines RNA with ATAC data, whereas the kScReNI variant only uses RNA data.

Feature selection. The inputs are the full RNA gene-count matrix and ATAC peak-count matrix. Genes and peaks are ranked by variability using Seurat v3 VST, after which only selected HVGs and highly variable peaks are retained for network inference [21]. The output is a reduced RNA matrix, a reduced ATAC matrix and the corresponding gene and peak lists. This step defines the feature space of the final GRNs, because genes removed here cannot appear as regulators or targets and peaks removed here cannot contribute to gene-peak-TF triplets or ATAC-derived weights.

Integration and cell clustering. The input is the selected scRNA-seq count matrix and selected scATAC-seq count matrix for the same cells. ScReNI applies Seurat’s weighted-nearest-neighbor (WNN) analysis, which independently reduces both modalities and then learns a per-cell weighting that reflects how informative each modality is for that cell [23]. The output is a shared WNN embedding with cluster assignments for all cells.

Neighbor identification. For each cell c , ScReNI identifies the K_{NN} nearest neighbors in the shared WNN embedding. The cell itself together with its K_{NN} neighbors defines the local population used to train the regulatory model for c . The output is a two-dimensional array of nearest-neighbor cell indices, with one row per sampled cell and K_{NN} columns listing the cells used as that cell’s local training neighborhood.

Gene-peak-TF association. Before any random forest is trained, ScReNI builds a restricted set of candidate regulatory links. First, ATAC peaks are linked to genes by genomic proximity, where only peaks within 250 kb upstream or downstream of a gene’s transcription start site are considered. Second, each candidate gene-peak pair with an absolute Spearman correlation below 0.1 across cells is removed. Third, TF binding motifs from TRANSFAC are matched to the retained peaks to identify which TFs have motif evidence for each gene-peak pair [24].

The result is a triplet table of the form (TF j , peak l , target gene i), encoding motif-supported TF-peak-target links that satisfy all proximity, correlation and motif criteria. This table is fixed for the remainder of the pipeline.

Random-forest inference. The random forest is the central model used by ScReNI to turn local single-cell measurements into regulatory edge weights. A random forest is an ensemble of decision trees, where each tree repeatedly splits the training samples according to feature values [25]. In this case the feature values include the expression of a candidate regulator or the accessibility of a chromatin peak to predict the target gene’s expression as accurately as possible. Rather than using the model’s output directly, ScReNI extracts feature importance scores from the random forest as evidence that an input feature helps explain the expression of a target gene.

Fitting a meaningful model from a single cell’s data would be statistically unreliable, so ScReNI assumes that nearby cells in the integrated single-cell space have similar regulatory programs. This assumption allows ScReNI to select the nearest neighbors of each cell and train a separate random forest for each target gene within this local neighborhood. The response vector y is the expression of the target gene measured in each neighbor cell. In kScReNI, the input matrix X contains the expression value of all other genes. Therefore, the importance of gene j is interpreted as the RNA-only regulatory weight from j to the target. In wScReNI, the input matrix is expanded to all ATAC peak-accessibility features in addition to all non-target genes, allowing the forest to assign importance to both gene-expression features and chromatin-accessibility features. The gene-peak-TF triplet table is later used to assign peak-feature importance scores to regulator-locus and target-peak terms.

Weight computation. The random forest returns a collection of target-specific feature importance vectors, which are converted to regulatory edge weights of the cell-specific GRN using the weight function. In kScReNI the forest only receives RNA-expression features, so the edge weight is simply the importance of the regulator’s expression feature when predicting the target:

$$k_{i,j} = d_{i,j}^k. \quad (1)$$

where $k_{i,j}$ denotes the edge weight from regulator gene j to the target gene i and $d_{i,j}^k$ is the importance assigned to gene j ’s expression feature.

In wScReNI the forest receives both RNA-expression features and ATAC peak-accessibility features. For each target i , the resulting feature importances are decomposed into three terms. The first is $d_{i,j}^w$, the importance of regulator j ’s expression feature in the wScReNI random forest. The second term is $r_{i,j}$, the summed importance of ATAC peaks assigned to regulator j ’s own locus. The target subscript i in $r_{i,j}$ indicates that these peak importances come from the random forest trained to predict target i . The third term is the motif-gated target-peak attribution. In the published implementation, this is the shared term $t_{i,j}$, which assigns chromatin importance from peaks near target i when regulator j has TF motif evidence for that target. The wScReNI weight formula derived from the R implementation can therefore be written as:

$$w_{i,j} = d_{i,j}^w + r_{i,j} + t_{i,j} \quad (2)$$

Let P_i be the set of ATAC peaks linked to target gene i and let $q_{i,l}^w$ be the importance assigned to peak l by the wScReNI random forest trained to predict target gene i . In the published

implementation, if regulator j appears in the set of TFs linked to target i , then it receives the sum of all retained peak importances linked to target i . Thus, every qualifying TF for the same target receives the same target-peak attribution, even if only a subset of those peaks contains motif evidence for that specific TF:

$$t_{i,j} = \begin{cases} \sum_{l \in P_i} q_{i,l}^w, & \text{if } I_{i,j} = 1 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

$$I_{i,j} = \begin{cases} 1, & \text{if } \exists l \in P_i \text{ with motif evidence for TF } j \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

The gene-peak-TF triplet table produced before network inference defines the binary indicator $I_{i,j}$, which marks whether at least one retained peak is linked to target gene i and contains motif evidence for TF j . The published shared term $t_{i,j}$ assigns the full target-linked peak-importance sum for P_i to the regulator j only when $I_{i,j} = 1$.

We additionally propose and evaluate a stricter TF-specific target-peak attribution. Instead of giving every qualifying TF the same total target-peak importance, this term sums only the target-associated peaks that contain motif evidence for regulator j :

$$t_{i,j}^* = \sum_{l \in P_{i,j}} q_{i,l}^w, \quad (5)$$

where $P_{i,j} \subseteq P_i$ is the set of peaks linked to the target i and motif-supported for TF j . If no such peak exists, $t_{i,j}^* = 0$. Since both target-peak attribution terms are computed from wScReNI peak-feature importances, an RNA-only kScReNI forest cannot produce them without ATAC peak predictors.

2.3 Formula variants

The main experimental question is how each part of the wScReNI weight formula contributes to the observed GRN changes relative to kScReNI. We evaluate six formula variants using top- k topology, ChIP-Atlas precision and network-based clustering ARI.

Variant	Formula	Purpose
$k_{i,j}$	$d_{i,j}^k$	default kScReNI
$w_{i,j}$	$d_{i,j}^w + r_{i,j} + t_{i,j}$	default wScReNI
$z_{i,j}$	$d_{i,j}^w + r_{i,j}$	regulator-locus wScReNI
$d_{i,j}^w$	$d_{i,j}^w$	ATAC-aware expression-only attribution
$s_{i,j}$	$d_{i,j}^w + t_{i,j}^*$	target-specific wScReNI
$v_{i,j}$	$d_{i,j}^k + t_{i,j}^*$	mixed-forest diagnostic variant

Table 2: Formula variants used to separate the contributions of random-forest feature space, regulator-locus accessibility and motif-gated target-peak attribution.

Instead of rerunning separate pipelines for every comparison, the final experiments decompose the saved network weights into reusable component matrices and recombine those components into formula variants. This keeps the cell sampling, gene set, peak set, random-forest outputs and metadata fixed while changing only the final edge-weight formula. In $v_{i,j}$, $t_{i,j}^*$ is still derived from the wScReNI peak-feature importances and only the expression-attribution base is taken from kScReNI.

2.4 Evaluation Metrics

The evaluation is designed to separate three questions that are often confounded in GRN benchmarking. First, we measure how much each formula component contributes to the final edge weights. Second, we quantify the extent to which each formula variant changes the top-ranked network topology relative to kScReNI. Third, we evaluate whether the top-ranked edges are biologically supported or preserve cell-type structure. All comparisons are made from the saved cell-specific weight matrices, so the formula variants differ only in how the same component matrices are combined, as described in Table 2.

Component composition. For each sampled cell, we sum the non-negative component matrices $d_{i,j}^w$, $r_{i,j}$ and $t_{i,j}$ over the complete gene-by-gene matrix. Each component is divided by the total weight of that cell ($d_{i,j}^w + r_{i,j} + t_{i,j}$), then the resulting percentages are averaged by cell type and across all sampled cells. This gives each cell equal weight in the summary, preventing cell types with more sampled cells from dominating the average. A second target-peak-specific analysis restricts the matrix to TF-target pairs where the theoretical motif indicator $I_{i,j} = 1$ and the per-cell target-peak attribution is positive. This evaluates the practical effect of $t_{i,j}^*$ on edges that can actually receive motif-gated ATAC support, rather than diluting the term over all possible gene pairs.

Top- k Jaccard similarity. Network topology is compared using the Jaccard similarity between the top- k edge sets. For two edge sets A and B ,

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

For each cell, we flatten the cell-specific weight matrix, select the 500 largest entries for each formula variant and compare that set with the top-500 set from $k_{i,j} = d_{i,j}^k$. This makes kScReNI the topology reference and evaluates whether each ATAC-informed formula preserves or replaces the RNA-only edge ranking. The same top-500 Jaccard procedure is also used to compare the R-compatible shared target-peak attribution with the stricter TF-specific target-peak attribution while holding the random-forest outputs fixed.

Chromatin immunoprecipitation sequencing (**ChIP-seq**) experiments identify DNA regions that bind to a particular TF, which provides more direct evidence of TF binding than motif matching. **ChIP-Atlas** is a public database of results from ChIP-seq experiments and was used in the ScReNI paper to evaluate whether the predicted MRD edges are supported by observed TF binding near the target gene. Although ChIP-Atlas is incomplete and should not be treated as ground truth, it can still be used as an external validation source for precision and recall.

ChIP-seq precision and recall. Edge-level ChIP-Atlas support is evaluated on MRD using ChIP-Atlas TF-target pairs overlapping the 500-gene network [26]. For every sampled cell and formula variant, the complete gene-by-gene matrix is ranked and the top-500 entries are selected, with self-edges retained to match the original ScReNI evaluation convention. An edge is counted as a true positive (TP) when its regulator-target gene pair appears in the ChIP-Atlas reference. Precision and recall are computed as

$$\text{precision} = \frac{\text{TP}}{500}, \quad \text{recall} = \frac{\text{TP}}{N_{\text{ChIP}}},$$

where N_{ChIP} is the number of ChIP-Atlas pairs whose TF and target both occur in the network gene set. Because every for-

mula is evaluated at the same top- k , precision is the main quantity reported in the results, while recall has the same variant ordering but on a different scale. ChIP-Atlas contains many public human ChIP-seq experiments, but these experiments are not evenly available across TFs, tissues, cell types and conditions [27; 28]. By contrast, SEA-AD validation would require TF-binding evidence matched to human middle temporal gyrus, cell subclass, disease state and experimental condition [29; 30]. Therefore, ChIP-Atlas is not used as SEA-AD ground truth for precision/recall analysis.

Network clustering ARI. As a complementary network-level check, we reproduce the degree-based clustering evaluation from the original ScReNI paper [2]. For each cell, the top- k entries of the weight matrix are selected and converted to a binary edge indicator matrix, with equal-weight ties resolved using R’s column-major matrix order to match the original ScReNI evaluation. Following the original evaluation convention, a gene-degree vector is computed from the column sums of this indicator matrix for each cell. Cells are clustered by complete-linkage hierarchical clustering using a distance matrix derived from correlations between log-transformed gene-degree profiles $\text{dist}(\text{cor}(\log(\text{degree} + 1)))$. The resulting clusters are compared with known retinal cell-type labels using the Adjusted Rand Index [31]. Across top- k thresholds from 400 to 1400, ARI measures how closely clusters formed from network-derived gene-degree profiles match the known retinal cell-type labels.

Literature reported microglia genes coverage. For the SEA-AD microglia experiment, the main evaluation occurs before GRN inference. The curated AD-related microglia regulator-target pairs are compared with the ranked microglia-specific HVG list. Because ranks are zero-based, genes with rank < 500 are considered present in the standard 500-HVG input. Genes with rank $< 2,000$ are considered present only after the expanded microglia-specific selection and genes outside the 2,000-HVG list are marked as absent. This check determines whether known AD regulatory relations are testable by the downstream GRN pipeline at all. If either the regulator or target is absent from the selected gene set, the corresponding edge cannot be inferred or ranked by ScReNI.

3 Results

3.1 Weight Formula Decomposition

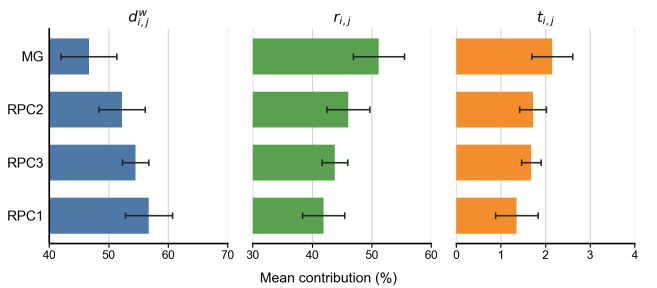


Figure 2: MRD decomposition of total wScReNI edge weight by cell type. Bars show the mean per-cell percentage contribution of the direct expression term $d_{i,j}^w$, the regulator-locus peak term $r_{i,j}$ and the motif-gated target-peak attribution $t_{i,j}$. The error bars show one standard deviation across sampled cells.

Using the component-composition metric defined in Section 2.4, we first measure how much each implemented wScReNI term contributes to the edge weights. The three reported terms are the direct expression importance $d_{i,j}^w$, the regulator-locus peak term $r_{i,j}$ and the motif-gated target-peak attribution $t_{i,j}$.

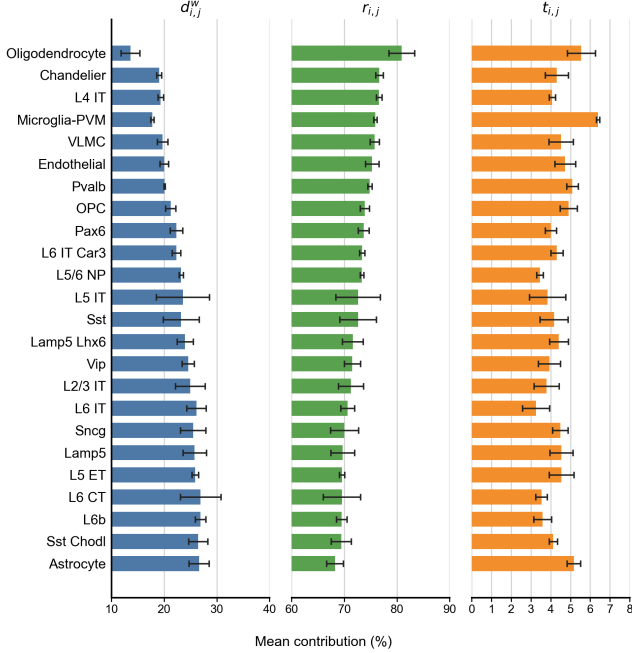


Figure 3: SEA-AD MTG decomposition of total wScReNI edge weight by subclass. Bars show the mean per-cell percentage contribution of $d_{i,j}^w$, $r_{i,j}$ and $t_{i,j}$. The error bars show one standard deviation across the three sampled cells per subclass.

The composition of components differs substantially between the datasets (Figures 2 and 3). In MRD, $d_{i,j}^w$ is the largest component overall at 52.5%, followed closely by $r_{i,j}$ at 45.7%. This balance is broadly consistent across retinal cell types, although $r_{i,j}$ is the largest component in MG cells (51.2%). In the sampled SEA-AD cells, $r_{i,j}$ dominates every subclass and contributes 72.8% overall, compared to 22.9% from $d_{i,j}^w$. Its subclass-level contribution ranges from 68.2% in astrocytes to 80.9% in oligodendrocytes. The target-peak attribution is the smallest component of total weight in both datasets, contributing 1.7% in MRD and 4.4% in SEA-AD.

3.2 Sparsity of Target-Peak Attribution

The small contribution of target-peak attribution to the total matrix weight follows from the definition of $I_{i,j}$, since $t_{i,j}$ is non-zero only for selected TF-target pairs. Its small share of total matrix weight does not imply that it has little influence where it is applied. The sparsity and edge-level impact of this term are examined in the following subsections.

We first measure the sparsity by finding the size of the candidate set that can receive target-peak attribution before analyzing its edge-level impact. For each dataset, we count all directed non-self gene-gene pairs among the 500 selected genes, then restrict this set to pairs where the regulator is annotated as a transcription factor and finally count the unique TF-target pairs

retained in the gene-peak-TF triplet table. These are the pairs for which $I_{i,j} = 1$.

Metric	MRD	SEA-AD
Genes	500	500
TFs	44	26
All gene-gene pairs	249,500	249,500
TF-regulator pairs	21,956	12,974
TF-regulator pairs (%)	8.80	5.20
$I_{i,j}=1$ pairs	5,769	5,298
$I_{i,j}=1$ pairs (%)	2.31	2.12
$I_{i,j}=1$ /TF pairs (%)	26.28	40.84
Triplet rows	9,954	10,424

Table 3: $I_{i,j}$ selection funnel for the MRD and SEA-AD datasets. Percentages for TF-regulator pairs and $I_{i,j}=1$ pairs are relative to all non-self gene-gene pairs. The $I_{i,j}=1$ /TF pairs percentage is relative to TF-regulator pairs only.

Table 3 shows that the target-peak attribution applies to only 2.31% of all candidate pairs in MRD and 2.12% in SEA-AD, because only TF regulators with retained target-associated motif evidence receive $I_{i,j}=1$.

3.3 Impact of $I_{i,j}$ Target-Peak Attribution

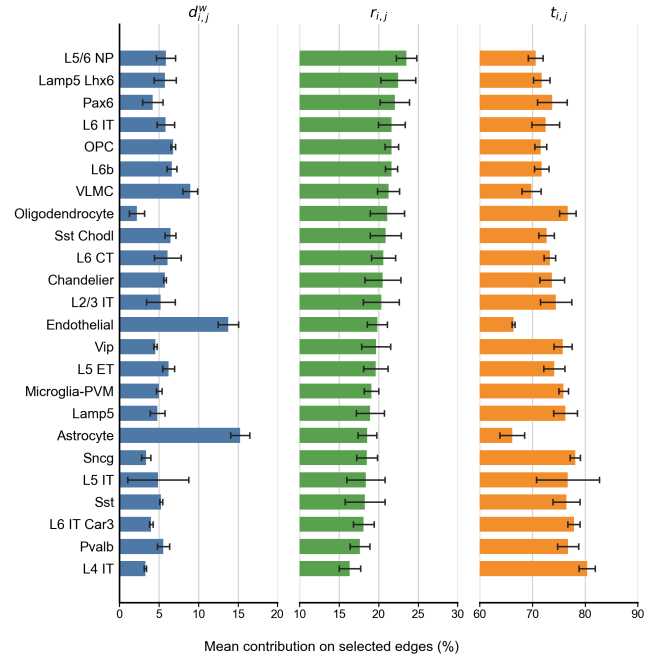


Figure 4: SEA-AD MTG component composition after restricting to edges with active target-peak attribution. Only edges with theoretical $I_{i,j}=1$ and positive per-cell target-peak attribution are included. The bars show mean per-cell percentages and error bars show one standard deviation across the three sampled cells per subclass.

The active-attribution analysis shows the opposite pattern from the global composition. After restricting to edges with $I_{i,j} = 1$ and positive per-cell $t_{i,j}$, the target-peak term is dominant: 80.6% of selected edge weight in MRD and 73.9% in SEA-AD. The mean number of active-attribution edges is 2,805 in MRD and 1,595 in SEA-AD. Thus, the term is sparse at the matrix level, but large on the subset of edges where it is actually applied.

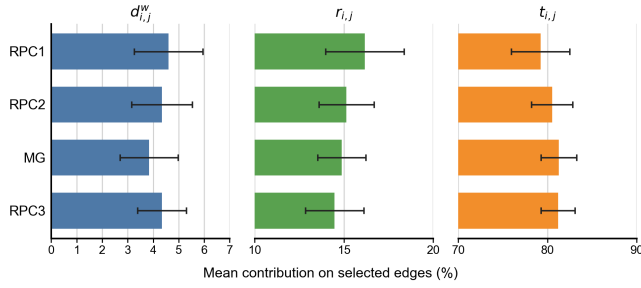


Figure 5: MRD component composition after restricting to edges with active target-peak attribution. Only edges with theoretical $I_{i,j}=1$ and positive per-cell target-peak attribution are included. The bars show mean per-cell percentages and error bars show one standard deviation across sampled cells.

3.4 Shared versus TF-Specific Target-Peak Attribution

We compare the original shared target-peak attribution with the stricter TF-specific target-peak attribution defined in Section 2. Because both networks use the same decomposed wScReNI random-forest output, this comparison isolates the target-peak attribution rule.

Cell type	Cells	Jaccard mean	Mean edges replaced
MG	10	0.9898	2.6
RPC1	10	0.9870	3.3
RPC2	10	0.9960	1.0
RPC3	10	0.9929	1.8
Overall	40	0.9914	2.18

Table 4: Top-500 edge-set similarity between the original shared target-peak attribution and the TF-specific alternative on the MRD dataset. Values are reported across sampled cells.

Table 4 shows that the two formulations produce nearly identical top-500 networks. The overall mean Jaccard similarity is 0.9914, corresponding to an average of 2.18 replaced edges per 500-edge network. Seventeen of the 40 sampled cells have identical top-500 edge sets. The largest mean change occurs in RPC1, but even there the average replacement is only 3.3 edges.

This result indicates that the shared target-peak attribution is not the main driver of the topology differences observed between kScReNI and wScReNI in this MRD benchmark. The TF-specific formulation is more precise because it assigns peak importance only to TFs with matching motif evidence at the contributing peaks, but its practical effect on the ranked networks is small at top-500. We therefore use this comparison as a sanity check on the target-peak attribution definition, rather than as evidence for a major network-level correction.

3.5 Jaccard Similarity of Formula Variants vs kScReNI

We compare the topology of each formula variant with the RNA-only kScReNI reference. For each sampled cell, we take the top-500 edges from each variant and compute their Jaccard similarity with the top-500 edges from $k_{i,j} = d_{i,j}^k$. This evaluates whether a variant preserves the kScReNI edge ranking while adding ATAC information, or instead replaces the RNA-derived network with a substantially different topology.

Figure 6 shows a consistent pattern across MRD and SEA-AD. The full $w_{i,j}$ network has very low overlap with kScReNI: mean Jaccard is 0.0076 in MRD and 0.0220 in SEA-AD. Removing the target-peak attribution but keeping $r_{i,j}$ gives essentially the same result ($z_{i,j}$: 0.0077 and 0.0229), showing that the regulator-locus peak term largely preserves the topology shift of the full formula. Removing $r_{i,j}$ increases overlap, especially in relative terms for SEA-AD, but the absolute overlap remains low. The target-specific variant $s_{i,j}$ reaches 0.0201 in MRD and 0.0752 in SEA-AD. The $d_{i,j}^w$ -only variant is similarly low (0.0214 and 0.0886), suggesting that the random forest itself changes the expression-importance ranking when ATAC features are included.

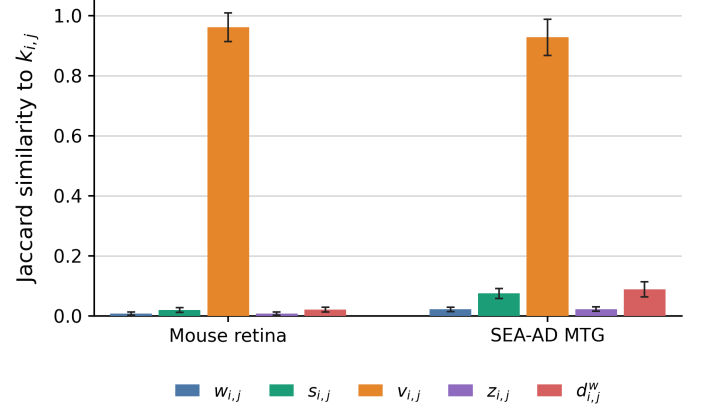


Figure 6: Top-500 edge-set Jaccard similarity between each formula variant and kScReNI on the MRD and SEA-AD MTG datasets. Error bars show one standard deviation across cells.

The only variant that remains close to kScReNI is $v_{i,j} = d_{i,j}^k + t_{i,j}^*$, with mean Jaccard 0.9610 in retina and 0.9278 in SEA-AD. This is unsurprising as the mixed-forest diagnostic keeps the original kScReNI expression weights and adds only the sparse wScReNI-derived target-peak attribution.

3.6 ChIP-Atlas Precision of Formula Variants

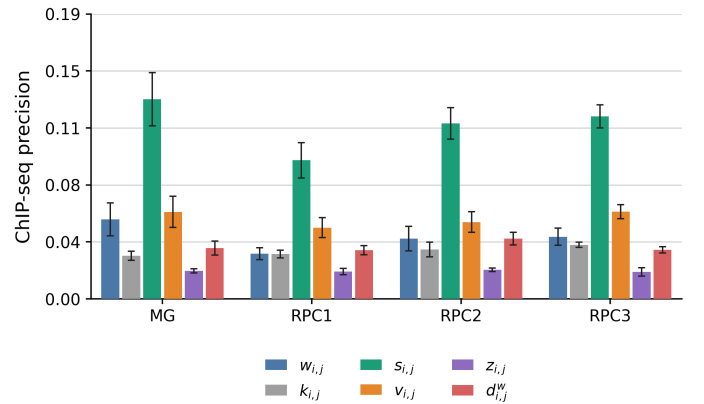


Figure 7: Mean ChIP-seq precision at the top-500 predicted edges for each network variant and retinal cell type ($n=10$ cells per type). Error bars show the standard error.

To test whether the topology changes correspond to improved biological support, we evaluate the MRD variants against ChIP-Atlas TF-target interactions using the top-500 precision/recall

metric. The benchmark contains 5,898 ChIP-Atlas pairs whose TF and target genes both occur in the 500-gene network and precision is shown by cell type in Figure 7.

The target-specific wScReNI formula $s_{i,j}$ gives the highest ChIP-Atlas precision at top-500 across all evaluated variants. Its overall mean precision is 0.1161, compared with 0.0412 for the published $w_{i,j}$ formula and 0.0318 for $k_{i,j}$. This is a 2.8-fold increase over the full wScReNI formula and a 3.6-fold increase over kScReNI. The same pattern is visible in every retinal cell type: $s_{i,j}$ reaches 0.1332 in MG, 0.0924 in RPC1, 0.1170 in RPC2 and 0.1216 in RPC3, while the published $w_{i,j}$ remains between 0.0300 and 0.0530.

The comparison between intermediate variants supports the decomposition interpretation. Within this MRD top-500 ChIP-Atlas benchmark, the $z_{i,j} = d_{i,j}^w + r_{i,j}$ variant has the lowest overall precision (0.0186), indicating that $r_{i,j}$ without target-peak attribution reduces TF–target precision relative to both $k_{i,j}$ and $d_{i,j}^w$. The $d_{i,j}^w$ -only variant is close to $k_{i,j}$ (0.0347 vs. 0.0318), while the mixed-forest variant $v_{i,j} = d_{i,j}^k + t_{i,j}^*$ improves to 0.0537 but remains below $s_{i,j}$. This suggests that the target-peak attribution is most effective when combined with the wScReNI expression attribution learned in the same random-forest feature space.

3.7 Clustering ARI of Formula Variants

As a complementary topology-level evaluation, we test whether each inferred network preserves cell-type-specific structure using the degree-based ARI metric from Section 2. We evaluate the same six variants as in the precision analysis over top- k thresholds from 400 to 1400.

The $d_{i,j}^w$ -only and $z_{i,j}$ variants achieve high ARI over parts of the threshold range 400–800, despite weaker ChIP-seq precision in the previous subsection. Thus, the ARI ranking is not identical to the ChIP-seq precision ranking.

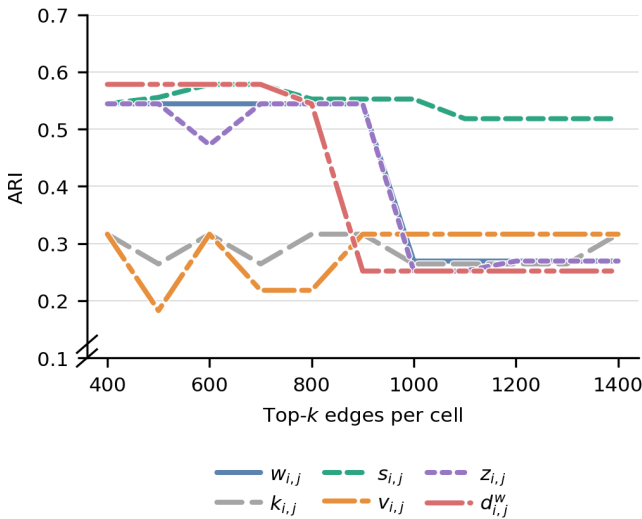


Figure 8: Network-based cell clustering ARI for the six evaluated formula variants at top- k edge thresholds from 400 to 1400.

Figure 8 shows that removing $r_{i,j}$ does not reduce cell-type separability in this evaluation. At top-500, $s_{i,j}$ obtains ARI 0.5555, slightly above the published-formula $w_{i,j}$ value of 0.5443 and clearly above $k_{i,j}$ at 0.2644. The target-specific formula reaches

its maximum ARI of 0.5783 at top-600 and top-700, matching the best value observed for the $d_{i,j}^w$ -only variant. In contrast, the mixed-forest diagnostic variant $v_{i,j}$ remains close to the kScReNI topology but does not improve clustering in this evaluation and reaches only 0.1828 in the top-500 and 0.3164 at most plotted thresholds.

3.8 Microglia Literature Pairs after HVG Selection

To test whether the SEA-AD microglia experiment could evaluate selected AD-related regulatory pairs, we compare a curated set of microglia-relevant literature pairs with the microglia-specific HVG ranks. Rank 0 denotes the most variable gene, rank < 500 corresponds to the standard top-500 HVG subset and genes absent from the 2,000-HVG list are reported as N/A.

Regulator	Target	Reg. rank	Target rank	Reference
TGFB1	P2RY12	N/A	N/A	[32]
TGFB1	TMEM119	N/A	N/A	[32]
RELA	IL1B	N/A	67	[32]
RELA	TNF	N/A	N/A	[32]
TREM2	APOE	N/A	1070	[33]
SPI1	TREM2	N/A	N/A	[34]
SPI1	TYROBP	N/A	N/A	[35]
SPI1	HLA-DRB1	N/A	N/A	[35]
SPI1	CSF1R	N/A	N/A	[35]
SPI1	AIF1	N/A	N/A	[35]
SPI1	IL6	N/A	N/A	[35]
SPI1	C3	N/A	N/A	[35]
SPI1	MMP9	N/A	N/A	[35]

Table 5: Microglia-relevant AD literature relations and their ranks in the microglia-specific SEA-AD HVG selection. N/A indicates absence from the 2,000-HVG set.

Table 5 shows that only two targets (IL1B and APOE) enter the expanded 2,000-HVG set, while none of the proposed regulators are retained. The standard 500-HVG setting would include even fewer genes, retaining only IL1B. Consequently, these literature relations cannot be ranked by ScReNI under this variance-based feature-selection, because at least one endpoint of each regulator-target pair is removed before network inference.

4 Discussion

The results show that the ATAC contribution in wScReNI should not be interpreted as a single, homogeneous signal. wScReNI combines the ATAC-aware expression importance $d_{i,j}^w$, the regulator-locus peak term $r_{i,j}$ and the target-peak attribution $t_{i,j}$. These components behave differently: $t_{i,j}$ is highly selective, while $r_{i,j}$ is applied independently of whether the regulator has motif evidence near the target. Treating the full formula as a single multi-omic correction therefore obscures which part of the chromatin information is target-specific.

This component separation supports an interpretation of why removing $r_{i,j}$ improves the MRD ChIP-Atlas precision while retaining the network-level signal captured by ARI. Accessible chromatin near a regulator’s own locus can indicate that the regulator gene is active or open for transcription factor binding, but it is not direct evidence that this regulator binds near a specific target. In a random forest, such features can still receive importance and change the ranking of other features, especially when predictors are correlated or redundant [36; 37]. The target-specific wScReNI formula $s_{i,j}$ therefore keeps

the ATAC-aware model fit and the motif-gated target-peak attribution, but removes the component most likely to add a non-specific accessibility signal.

The topology results show that preserving the kScReNI edge ranking and improving TF-target support are different goals. The mixed-forest diagnostic variant $v_{i,j}$ stays close to kScReNI because it starts from the RNA-only weights and adds only the sparse wScReNI-derived target-peak attribution. The target-specific formula $s_{i,j}$ instead remains a wScReNI variant: it inherits the change in feature importance ranking caused by adding ATAC features to the predictor matrix. This distinction matters because a network can be close to the RNA-only topology without being the best-supported network under ChIP-Atlas, while a more biologically supported network can still be topologically distant from kScReNI.

Some formula variants score well on ARI despite lower ChIP-Atlas precision, which is unsurprising because ARI does not measure whether individual TF-target edges are correct. Specifically, it evaluates whether cells with the same MRD cell-type label have similar network-derived gene-degree profiles, following the network-based clustering evaluation in the original ScReNI paper [2]. A formula can therefore preserve broad cell-type differences in expression or accessibility while still assigning top-ranked TF-target edges with weak ChIP-Atlas support.

The SEA-AD analysis should be interpreted more conservatively than the MRD benchmark. SEA-AD is useful for testing whether the same component-level behavior appears in human cortex, but it is not used as a ChIP-seq-validated precision benchmark for the validation limitations described in Section 2.4. The SEA-AD results therefore support component-pattern replication and scope characterization, rather than validating $s_{i,j}$ for human AD cortex.

The microglia-specific HVG experiment reveals an additional limitation for AD applications. Even after expanding selection to 2,000 HVGs and retaining all 3,129 Microglia-PVM cells, the selected gene set still excludes every regulator in the curated microglia AD literature relations. This is most likely because variance-based feature selection prioritizes genes that vary across cells, while some biologically central regulators may be relatively stable in expression variance within the Microglia-PVM cell type. For direct AD hypothesis testing, ScReNI therefore needs a feature-selection strategy that explicitly retains known TFs and disease-relevant regulators instead of relying only on variance ranking.

Several limitations remain. First, the strongest biological validation is restricted to the MRD benchmark, where ChIP-Atlas overlap is available for the selected genes. The same precision gain has not yet been confirmed in a matched human brain benchmark. Second, the ChIP-Atlas caveats described in Section 2.4 still apply to the MRD precision results. Finally, the SEA-AD microglia result shows that the pipeline can fail before network inference begins if feature selection removes relevant regulators. This limitation applies not only to ScReNI, but to any GRN workflow that starts from a variance-ranked gene set.

5 Conclusions and Future Work

This work asked how the ATAC-derived terms in wScReNI change cell-specific GRNs and whether those changes make the method more useful for studying Alzheimer’s disease regulatory mechanisms. The results reveal both a limitation of the published formula and a promising refinement. A

target-specific use of ATAC-derived peak importance improves ChIP-Atlas precision at top-500 among the tested variants in the mouse retina development benchmark, but the published wScReNI formula does not function as a minor chromatin adjustment to the RNA-only kScReNI network. In both MRD and SEA-AD, the full wScReNI ranking has low top-500 overlap with the RNA-only ranking, while the mixed-forest diagnostic variant $v_{i,j} = d_{i,j}^k + t_{i,j}^*$ stays close to kScReNI.

The main methodological conclusion is that the regulator-locus peak term $r_{i,j}$ should not be treated as interchangeable with target-specific chromatin evidence. After separating the terms implemented in ScReNI, the mouse retina development validation supports the target-specific wScReNI formula $s_{i,j} = d_{i,j}^w + t_{i,j}^*$.

At top-500, this variant reaches mean ChIP-Atlas precision 0.1161, compared with 0.0412 for the published $w_{i,j}$ formula and 0.0318 for kScReNI. It also keeps the network-based cell-type signal used in the original ScReNI evaluation: its top-500 ARI is 0.5555, close to the published formula’s 0.5443 and above the RNA-only baseline’s 0.2644 in this sample.

For SEA-AD, the conclusion is more limited. The same components can be decomposed in human cortex data and the full wScReNI formula again produces a network that is far from the RNA-only baseline. However, this dataset was not used as a matched TF-binding validation benchmark. The microglia-specific experiment also shows that the current pipeline cannot directly test several known AD regulatory hypotheses, as all checked regulator genes are absent in the selected 2,000 microglia-specific HVGs from the full set of 3,129 Microglia-PVM cells. The filtering step happens before network inference, so changing the downstream wScReNI formula cannot recover those hypotheses on its own.

5.1 Further Research

Future work should first validate $s_{i,j}$ on datasets with matched TF-binding evidence beyond MRD. For Alzheimer’s disease, this requires human, brain-region-matched and preferably cell-type-specific ChIP-seq data. Generic all-human ChIP-seq overlap would not be sufficient for a strong claim because TF binding depends on tissue, cell type and condition.

A second direction is to adapt gene selection for disease regulatory analysis. Variance-ranked HVG selection is useful for reducing the feature space, but it can remove stable regulators that are central to the biological question. A more appropriate SEA-AD pipeline would combine HVG ranking with forced inclusion of literature-supported TFs, TRANSFAC-registered TFs and disease-relevant targets (Table 5), then complete the feature set with VST-ranked highly variable genes.

After such feature selection, $s_{i,j}$ should be tested in differential GRN analyses across disease stage, donor metadata and cell subclass. Another direction is to evaluate whether the same component behavior holds for ScReNI’s unpaired RNA and ATAC setting, which was outside the scope of this paired-multiole analysis. Finally, the TF-specific target-peak attribution should be evaluated on datasets with denser ATAC coverage. In the current mouse retina development sample it changes few top-500 edges relative to the shared target-peak attribution, but it is conceptually more specific and may become more important when more peaks satisfy the gene-peak-TF filtering criteria.

6 Responsible Research

6.1 Reproducibility

The full code repository is available at <https://github.com/DelftBioinformaticsLab/bsc-screni>. The ScReNI reimplementation and analysis scripts are written in Python using publicly available libraries. The repository also includes SLURM scripts to run computationally intensive steps on an HPC cluster. The SEA-AD and mouse retina development data used in this work are publicly released by their original data-generating studies and remain subject to the access conditions of those sources.

Randomness is controlled explicitly and the seeds are included in the scripts. Cell subsampling uses fixed seeds and the random-forest models are run with a fixed `random_state` parameter, which controls the estimator’s random seed in the Python scikit-learn implementation. The final retinal decomposition uses 40 sampled cells, with 10 cells per retinal cell type and the SEA-AD decomposition uses 72 sampled cells, with three cells per subclass. The evaluation scripts also document their ranking conventions, including top-500 edge selection and the column-major top- k ranking used to reproduce the original ScReNI ARI evaluation. These details are important because small changes in edge-ranking conventions can change top- k network metrics.

6.2 Data and Ground Truth

This work uses previously released research datasets. The MRD data are used as a benchmark because they match the original ScReNI evaluation setting [2]. SEA-AD is used as a human Alzheimer’s disease context because it provides single-nucleus RNA and ATAC data from middle temporal gyrus across disease progression [22]. No new human samples are collected in this work.

The strongest external validation in the paper is restricted to the MRD dataset. As described in Section 2.4, ChIP-Atlas is used as an external support set for predicted MRD TF-target edges, not as complete biological ground truth. Absence from ChIP-Atlas therefore does not prove that a predicted edge is false.

For SEA-AD, the same validation boundary applies. ChIP-Atlas is not used as a precision benchmark, so the SEA-AD results should be read as mechanism replication and scope characterization, not as a validated human AD GRN benchmark.

The gene-peak-TF triplets depend on motif evidence, including motifs from TRANSFAC [24]. Only derived triplet counts, edge weights and summary statistics are reported in this paper. The licensed motif database itself is not redistributed.

6.3 Limitations and Potential Misuse

The main misuse risk is overinterpreting inferred GRN edges as experimentally confirmed regulation. ScReNI produces weighted predictions from single-cell expression, chromatin accessibility, motif evidence and random forest feature importance. These signals can suggest regulatory hypotheses, but they do not establish causality. The target-specific wScReNI formula $s_{i,j}$ evaluated in this paper is assessed using MRD ChIP-Atlas precision and ARI analyses. However, it is not clinically validated and should not be used to make diagnostic, prognostic, or treatment decisions.

A second risk is overstating the AD relevance of the SEA-AD analyses. The SEA-AD experiments show that the same formula components can be decomposed in human cortex data

and that the current HVG-based pipeline fails to include many canonical microglia AD regulators. They do not prove that $s_{i,j}$ identifies AD regulatory mechanisms in human brain. In particular, the microglia-specific experiment shows that known literature relations are often untestable because the regulator genes are removed before network inference. Reporting downstream network rankings for those relations without stating this feature-selection boundary would be misleading.

The evaluation metrics also have limited scope. ChIP-seq precision evaluates TF-target support for top-ranked edges in MRD, while ARI evaluates whether network-derived degree profiles preserve cell-type labels. A formula can score well on ARI by preserving broad cell-type structure without improving edge-level regulatory accuracy. Conversely, a formula with better TF-target precision may still have low top- k overlap with the RNA-only network. The paper therefore reports topology, precision and clustering metrics separately instead of treating any single metric as definitive.

6.4 Ethical Considerations

This study is a secondary computational analysis of de-identified public research data. Ethical approval and consent for donor tissue collection fall under the original data-generation studies. The present work does not attempt to re-identify donors and does not use individual-level results for clinical interpretation.

The social and clinical relevance of Alzheimer’s disease research makes cautious and precise reporting especially important. The paper therefore separates validated MRD findings from unvalidated human AD cortex analyses and explicitly states when a regulatory relation cannot be tested because the relevant genes are absent from the feature set. This reduces the risk that computational network predictions are presented as established disease mechanisms. Future use of this pipeline on human AD data should include matched validation data and feature-selection strategies that retain known disease-relevant regulators before making biological claims.

6.5 Use of Generative AI

Generative AI was used as a supporting tool while working on this project. Early in the implementation work, it was used to help debug Python code and to suggest possible ways of translating parts of the original ScReNI R implementation into Python. Generative AI was also used to improve documentation and to add clearer debug information to output logs, to make it easier for future users to understand and use.

In addition to assistance with correcting spelling and grammar, it was also used to suggest more formal academic phrasing during writing. For Figure 1, AI assisted with the creation of vector-based elements used in the image.

For biology-related outputs, AI assistance was used only to support interpretation and to identify points that required further examination. It was not treated as biological evidence, all scientific claims in this paper were based on the reported analyses and cited literature.

A ScReNI Workflow Schematic

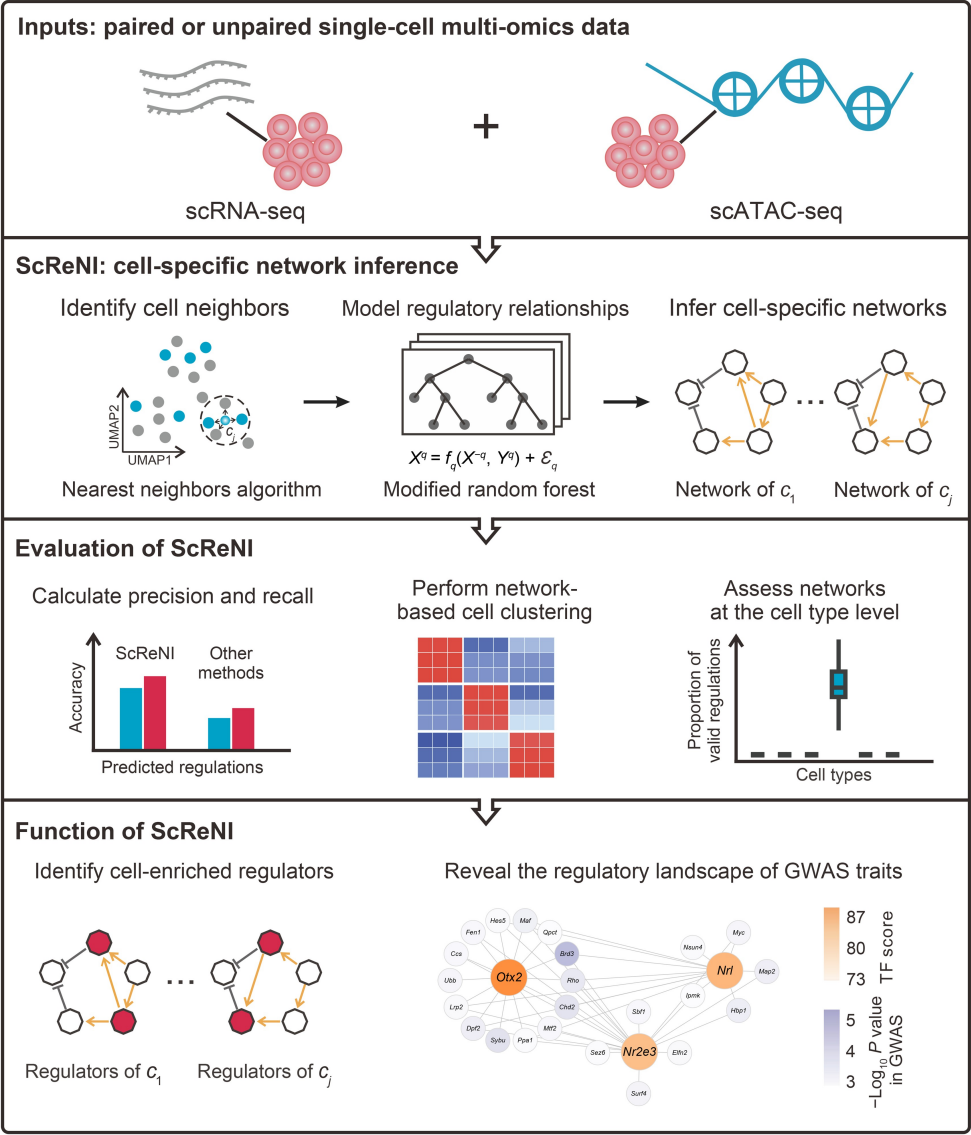


Figure 9: Overview of the original ScReNI workflow, including single-cell multi-omics input, cell-specific network inference, evaluation and downstream regulator analysis [2].

References

- [1] World Health Organization. Dementia. <https://www.who.int/news-room/fact-sheets/detail/dementia>, March 2025. Fact sheet, accessed 2 June 2026.
- [2] Xueli Xu, Yanran Liang, Miaoxiu Tang, Jiongliang Wang, Xi Wang, Yixue Li, and Jie Wang. ScReNI: Single-cell regulatory network inference through integrating scRNA-seq and scATAC-seq data. *Genomics, Proteomics & Bioinformatics*, 23(4):qzaf060, 2025.
- [3] Michael A. DeTure and Dennis W. Dickson. The neuropathological diagnosis of Alzheimer’s disease. *Molecular Neurodegeneration*, 14(1):32, 2019.
- [4] Christopher H. van Dyck, Chad J. Swanson, Paul Aisen, et al. Lecanemab in early Alzheimer’s disease. *The New England Journal of Medicine*, 388(1):9–21, 2023.
- [5] John R. Sims, J. Anthony Zimmer, Cynthia D. Evans, et al. Donanemab in early symptomatic Alzheimer disease: the TRAILBLAZER-ALZ 2 randomized clinical trial. *JAMA*, 330(6):512–527, 2023.
- [6] Vivek Swarup, Flora I. Hinz, Jessica E. Rexach, Ken-ichi Noguchi, Hiroyoshi Toyoshiba, Akira Oda, Keisuke Hirai, Arjun Sarkar, et al. Identification of evolutionarily conserved gene networks mediating neurodegenerative dementia. *Nature Medicine*, 25(1):152–164, 2019.
- [7] Ryan A. Neff, Minghui Wang, Sezen Vatansever, Lei Guo, Chen Ming, Qian Wang, Erming Wang, Emrin Horgusluoglu-Moloch, et al. Molecular subtyping of Alzheimer’s disease using RNA sequencing data reveals novel mechanisms and targets. *Science Advances*, 7(2):eabb5398, 2021.
- [8] Hansruedi Mathys, Jose Davila-Velderrain, Zhuyu Peng, et al. Single-cell transcriptomic analysis of Alzheimer’s disease. *Nature*, 570(7761):332–337, 2019.
- [9] Yingyue Zhou, Wilbur M. Song, Prabhakar S. Andhey, Amanda Swain, Tyler Levy, Kelly R. Miller, Pietro L. Poliani, Manuela Cominelli, et al. Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer’s disease. *Nature Medicine*, 26(1):131–142, 2020.
- [10] Alexi Nott, Inge R. Holtman, Nicole G. Coufal, Johannes C. M. Schlachetzki, Miao Yu, Rong Hu, Claudia Z. Han, Monique Pena, et al. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science*, 366(6469):1134–1139, 2019.
- [11] M. Ryan Corces, Anna Shcherbina, Soumya Kundu, Michael J. Gloudemans, Laure Frésard, Jeffrey M. Granja, Bryan H. Louie, Tiffany Eulalio, et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer’s and Parkinson’s diseases. *Nature Genetics*, 52(11):1158–1168, 2020.
- [12] Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16:133–145, 2015.
- [13] Pau Badia-I-Mompel, Lodewyk Wessels, Sophia Müller-Dott, Rémi Trimbou, Ricardo O. Ramirez Flores, Ricard Argelaguet, and Julio Saez-Rodriguez. Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics*, 24:739–754, 2023.
- [14] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10:1213–1218, 2013.
- [15] Carmen Bravo González-Blas, Sara De Winter, Gert Hulselmans, Nikolai Hecker, Iulian Matetovici, Valerie Christiaens, et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature Methods*, 20:1355–1367, 2023.
- [16] Jonas Simon Fleck, Sophie Martina Johanna Jansen, Damian Wollny, Fides Zenk, Makiko Seimiya, Ankit Jain, Ryohei Okamoto, Malte Santel, Zhisong He, J. Gray Camp, and Barbara Treutlein. Inferring and perturbing cell fate regulomes in human brain organoids. *Nature*, 621(7978):365–372, 2023.
- [17] Junyue Cao, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361:1380–1385, 2018.
- [18] Sai Ma, Bing Zhang, Lindsay M. LaFave, Andrew S. Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Amanda Brack, Vinay K. Kartha, Timothy Tay, Travis Law, Caleb A. Lareau, Yu-Chieh Hsu, Aviv Regev, and Jason D. Buenrostro. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell*, 183(4):1103–1116.e20, 2020.
- [19] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.
- [20] Peng Lyu, T. Hoang, C. P. Santiago, E. D. Thomas, A. E. Timms, H. Appel, et al. Gene regulatory networks controlling temporal patterning, neurogenesis, and cell-fate specification in mammalian retina. *Cell Reports*, 37:109994, 2021.
- [21] Tim Stuart, Andrew Butler, Paul Hoffman, et al. Comprehensive integration of single-cell data. *Cell*, 177:1888–1902, 2019.
- [22] Mariano I Gabitto, Kyle J Travaglini, Victoria M Rachleff, et al. Integrated multimodal cell atlas of Alzheimer’s disease. *Nature Neuroscience*, 27:2366–2383, 2024.
- [23] Yuhao Hao, Stephanie Hao, Erica Andersen-Nissen, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, 2021.
- [24] V Matys, OV Kel-Margoulis, E Fricke, et al. TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34:D108–D110, 2006.
- [25] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [26] Shinya Oki, Tazro Ohta, Go Shioi, et al. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Reports*, 19(12):e46255, 2018. ChIP-Atlas ground truth database.

- [27] Zhen Zou, Tazro Ohta, and Shinya Oki. ChIP-Atlas 3.0: a data-mining suite to explore chromosome architecture together with large-scale regulome data. *Nucleic Acids Research*, 52(W1):W45–W53, 2024.
- [28] Saeko Tahara and Haruka Ozaki. Unmeasured human transcription factor ChIP-seq data shape functional genomics and demand strategic prioritization. *Briefings in Functional Genomics*, 24:elaf016, 2025.
- [29] Jacob M. Loupe, Ashlyn G. Anderson, Lindsay F. Rizzardi, Ivan Rodriguez-Nunez, Belle Moyers, et al. Extensive profiling of transcription factors in postmortem brains defines genomic occupancy in disease-relevant cell types and links TF activities to neuropsychiatric disorders. *bioRxiv*, 2023. Preprint.
- [30] Ashlyn G. Anderson, Belle A. Moyers, Jacob M. Loupe, et al. Allele-specific transcription factor binding across human brain regions offers mechanistic insight into eQTLs. *Genome Research*, 34(8):1224–1234, 2024.
- [31] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.
- [32] Oleg Butovsky, Mark P Jedrychowski, Craig S Moore, et al. Identification of a unique TGF- β -dependent molecular and functional signature in microglia. *Nature Neuroscience*, 17(1):131–143, 2014.
- [33] Susanne Krasemann, Charlotte Madore, Ron Cialic, et al. The TREM2–APOE pathway drives the transcriptional phenotype of dysfunctional microglia in neurodegenerative diseases. *Immunity*, 47(3):566–581, 2017.
- [34] Zhen Wei, Xiaodong Pan, Xiaoli Cui, Jing Zhang, Xiaoman Dai, Yuqi Zeng, and Xiaochun Chen. PU.1 dictates beta-amyloid-induced TREM2 expression upregulation in microglia in a transgenic model of Alzheimer’s disease. *Frontiers in Aging Neuroscience*, 17:1537388, 2025.
- [35] Justin Rustenhoven, Delphine Jansson, Leon C Smyth, and Mike Dragunow. PU.1 regulates Alzheimer’s disease-associated genes in primary human microglia. *Molecular Neurodegeneration*, 13:44, 2018.
- [36] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, 8:25, 2007.
- [37] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307, 2008.