Decoding Gene Regulation in Alzheimer's disease with Transfer Learning and Explainable Machine Learning

Sergio Peignier, Amanda Lo Van, Yann Meunier, Elea Pauliat, Matis Zouari, Federica Calevro

Decoding Gene Regulation in Alzheimer's disease with Transfer Learning and Explainable Machine Learning

Sergio Peignier[1,†][0000−0002−9004−3033], Amanda Lo Van[2,*], Yann Meunier[1], Elea Pauliat[1], Matis Zouari[1], and Federica Calevro[3][0000−0001−7856−9617]

[1] INSA Lyon, INRAE, BF2I, UMR0203, F-69621, Villeurbanne, France
[2] INSA Lyon, CNRS, LaMCoS, UMR5259, 69621 Villeurbanne, France
[3] INRAE, INSA Lyon, BF2I, UMR0203, F-69621, Villeurbanne, France
[†] sergio.peignier@insa-lyon.fr
[*] amanda.lo-van@insa-lyon.fr

Abstract. Gene Regulatory Networks are major mechanisms governing biological processes. These networks are often inferred from gene expression matrices using Machine Learning algorithms. Nevertheless this task requires large datasets, to avoid learning unsuitable links due to over-fitting. This limitation impedes applying safely such tools on gene expression datasets with limited number of samples, such as rare cell-types or diseases. Moreover, the intrinsic black-box nature of most Machine Learning tools is an important drawback in sensitive areas such as in the bio-medical context. In this work we propose a new method, based on Transfer Learning and Semi-Supervised Learning, to infer interpretable gene regulatory networks, when only few samples are available. We have applied this tool to infer the first gene regulatory network of cell-line SH-SY5Y, a well recognized *in vitro* model for Alzheimer's disease. This biomedical application is particularly important, since Alzheimer's disease is a poorly understood neurodegenerative disease that causes around 60-70% of dementia cases worldwide and since there are currently no treatments to stop or reverse it. Our transfer methodology was successfully assessed qualitatively, and we revealed that our simple and interpretable models remained competitive w.r.t state-of-the-art complex models. The exploration of the regulatory links, revealed that our methodology identified well-known regulatory pathways involved in Alzheimer's disease. The encouraging outcomes yielded by our approach suggest its potential for analyzing other small gene expression datasets.

Keywords: Transfer Learning · Explainable Machine Learning · Gene Regulatory Network · Alzheimer's Disease.

1 Introduction

Alzheimer's Disease (AD) is a neurodegenerative disorder that accounts for 60-70% of dementia cases in humans, affecting more than 55 million individuals

worldwide [16]. It is characterized by cerebral atrophy and a decrease in synaptic transmissions, involving neural death and the loss of inter-neuron connections in the brain [5]. Currently there exists no real cure for AD patients and current treatments mostly focus on altering the biological pathways that are involved in the development of symptoms [45]. AD involves the accumulation of plaques of repeating units of an organic compound called Amyloid Beta ($A\beta$), due to the erroneous cleavage of the so-called Amyloid Precursor Protein (APP). The accumulation of $A\beta$ oligomers leads to the generation of Reactive Oxygen Species (ROS), i.e., compounds that can easily react with other molecules in a cell, thus increasing the cellular oxidative stress, which induces an activation of immune responses and defense mechanisms [44]. In addition to studying post-mortem samples from AD patients, and in order to study functionally the AD process, researchers have relied both on animal and cell-culture models. Compared to *in vivo* animal models, *in vitro* cell culture, such as the cell line SH-SY5Y, provides a robust, efficient and cost-saving method to study the pathology of specific diseases, with a possible focus on the characterization of molecular mechanisms [22].

Gene Regulatory Networks (GRNs) have been recognized as major mechanisms controlling key biological processes [17]. The development of high-throughput sequencing techniques has allowed to quantify gene expression in different conditions, tissues and time-points (using bulk RNA-seq and microarray experiments) and even in individual cells (using single-cell RNA-seq). Such massive data acquisition has enabled the development of plethora of computational tools to infer gene regulatory interactions from gene expression data, as reported in [12,20]. Among the different families of inference methods, the so-called data-driven paradigm is recognized as efficient and accurate [12]. Data-driven methods use gene expression data to score potential links between regulators and genes, and then select the most promising links. Earlier approaches relied on correlation or mutual information metrics to score the relatedness between genes and form co-expression networks. Most recent data-driven methods aim at training regression/classification algorithms to model the expression of each gene as a function of regulatory genes' expressions. Then, they use a feature importance scoring procedure to identify for each gene, the subset of regulators that contribute the most to the predictive task (e.g., [30,1]).

However, inferring GRNs in a data-driven manner requires large gene expression matrices in order to prevent the inference tools from over-fitting and learning unsuitable links. This limitation impedes applying safely such tools on gene expression datasets with limited number of replicates, e.g. on non-model organisms, rare cell-lines or rare diseases. For instance, in the context of AD, scarce gene expression data are available for the SH-SY5Y cell line, which has prevented the use of traditional data-driven tools to infer the regulatory interactions governing such cell line. Interestingly, in the ML domain, the methodology termed Transfer Learning (TL), aims at taking advantage of data obtained in different domains, to cope with the lack of data [42]. TL has been used successfully in different real-world applications and thus has gained increasing attention in the

ML community [42]. However, despite its importance, TL has been marginally used in the bio informatics field in general [23], and rather few TL techniques have been applied to GRN inference.

In addition, most ML models share another important limitation: their outputs and predictions' lack of interpretability. Indeed, ML models' large number of parameters and complexity make them inherently unintelligible for human beings. The intrinsic black-box nature of such systems could be an important drawback in sensitive areas such as in the inference of GRN in bio-medical contexts. According to [26], in the ML domain, in order to overcome such a lack of intepretability, authors have focused on two major methodologies: i) using interpretable models, including linear models or decision trees, or ii) using interpretation tools, that can be applied a posteriori, in order to assess the importance of each feature in the predictive task, in a global or in a local manner, i.e., for the entire dataset or for each specific instance. Despite the importance of interpretability in bio-medical applications, most data-driven GRN inference tools rely on complex black-box models, and global feature importance tools have been used to infer regulatory links, and the use of interpretable ML models has been under-explored.

In this work, we propose a simple unsupervised TL tool for RNA-seq data, based on a linear feature-based domain adaptation method and a semi-supervised label spreading algorithm, in order to train explainable decision-trees (DTs) and build interpretable GRNs for datasets with rather few samples. Moreover, as a study case, we have applied this new technique to infer an interpretable GRN of the SH-SY5Y cell-line model. Given the absence of prior research in this area and the significance of SH-SY5Y as a model for studying neurodegenerative diseases, including AD, this application holds particular promise for advancing AD research *in vitro*. However, the proposed method is not exclusive to AD and could be applicable to the analysis of other small datasets.

The rest of the paper is organized as follows: In Section 2 we describe previous works on GRN inference in AD and the application of TL to the inference of GRNs. In Section 3 we present our methodology. In Section 4 we present the results, and in Section 5 we conclude with a discussion on the results and some perspectives.

2 State of the art

2.1 Transfer Learning applied to GRN inference

An important premise in ML states that both the training and the test datasets should belong to the same domain, i.e., they should share the same distributions in the feature space. Nevertheless, in some real world applications, this assumption may not hold, and collecting new data to train a model in the application domain may reveal to be difficult and expensive. A recent methodology, termed Transfer Learning (TL), aims at taking advantage of data obtained in different domains, to cope with the aforementioned problem [42]. This methodology has

been used successfully in different real-world applications and thus has gained increasing attention in the machine learning community [42].

Despite its importance, TL has been marginally used in the bio-informatics field in general [23], and rather few TL techniques have been applied to GRN inference, as reported hereafter. In [23], authors proposed a pioneer methodology to reconstruct the GRN of a target organism, by exploiting available information in a source organism using supervised TL. GRN inference was framed as a binary classification task, where potential regulatory links were labeled as *True* or *False*, depending on the expression profiles of the genes involved in the link. In practice, given the lack of negative examples, only regulatory links experimentally validated were used as positive examples, whereas other links were considered as unlabeled ones. Then, the authors relied on a semi-supervised weighted-SVM to determine positive links. Nevertheless, such a semi-supervised tool cannot be applied when positive examples in a specific source context are not available. In [37], a parameter-based TL methodology for GRN inference has been presented. In this work, an attention-based deep learning model was pre-trained, in a self-supervised manner, on a large source dataset containing around 30 million single-cell RNA-seq libraries. Then, the authors fine-tuned the aforementioned network on different specific tasks with limited target data. Similar methodologies based on deep neural networks, pre-trained on very large source databases, have also been proposed in other works, e.g., [10,9]. Nevertheless, such methods require a pre-training stage on a very large single-cell RNA-seq source dataset from a given species, and are not directly applicable to non-model species/condition, for which only smaller bulk RNA-seq or micro-array datasets are available. Moreover, given the number of parameters and the black-box nature of deep learning models, it is hard to determine whenever the system incurs on negative-transfer. Indeed, the ability of such networks to generalize to data that differs from the training data distribution, is still being explored, since inaccurate behaviours, have been reported for instance in image analysis (e.g., [3]) and natural language processing (e.g., [14]), which could cause important concerns in sensitive domains such as in bio-medicine. Moreover, the aforementioned techniques share another important limitation: their outputs and predictions' lack of interpretability. In [25], interpretability is defined as the "degree to which a human can understand the cause of a decision". Given the large number of parameters, and their complexity, these systems are inherently unintelligible for human beings. The intrinsic black-box nature of such systems could be an important drawback in areas such as the bio-medical one.

2.2 Gene Regulatory Inference in AD

Given the importance of AD, previous works have addressed the inference of underlying GRNs involved in this pathology. Shared co-expression modules between AD and epilepsy [40], and between AD and severe Covid-19 [15] have been identified from human RNA-seq data. In both studies, further analysis revealed that such co-expression modules exhibit an important enrichment for genes involved in nervous system-related pathways. A generalization-aware self-expressive GRN

inference method, proposed in [28], was applied to human RNA-seq data to identify regulatory modules over-expressed in AD. Such modules revealed to be enriched with genes involved in immune response. GRNs involved in molecular processes taking place before the detection of neuropathology, were inferred in transgenic male and female AD mice, from neocortex single-cell RNA-seq data [2]. This study revealed key gender differences associated to the disease, mainly related to ROS metabolism and synapses organization. In [31], the authors proposed a framework to infer large GRNs in parallel, using 15 state-of-the-art methods. The authors tested their framework using in silico benchmarks and real-world gene expression matrices, including a gene expression dataset from a transgenic AD mouse model. Then, the resulting GRN was clustered into regulatory modules, several of which were enriched with genes responsible for a variety of brain disorders. A GRN inference approach based on gradient boosting and mutual information was used to detect GRNs from both: single-cell RNA-seq data from prefrontal cortex samples from control and AD patients, and spatial RNA-seq data from tissue regions around Aβ aggregates in AD mouse model [18]. This study identified regulatory modules enriched with pathways involved in AD such as immune and inflammatory processes.

3 Materials and Methods

3.1 Data processing

In order to infer the GRN involved in AD on SH-SY5Y cells, we have collected different datasets from the *gene expression atlas*[4]: 1) **Allen Brain Atlas** [24], a bulk RNAseq dataset reporting the expression of human genes in 377 samples from the white matter of forebrain, the hippocampus and the neocortex of *control* and *AD elderly individuals*. 2) **Hisayama study** [11], a microarray dataset reporting the expression of human genes in 161 libraries from neurons belonging to the cortex, the hippocampus and the gyrus of *control* and *AD elderly individuals*. 3) **PS2APP** [33], a bulk RNAseq dataset reporting the expression of mouse genes in 61 libraries from astrocytes, microglia and neurons, sampled from the cerebral cortex of *control* and *PS2APP transgenic mice*, exhibiting an AD-like pathology, with important cerebral Aβ levels and neuroinflammation. 4) **SH-SY5Y** [39], a microarray dataset containing the expression of human genes in 12 libraries corresponding to SH-SY5Y cell-lines transfected to express low, medium or high levels of Aβ, and thus simulate AD-like phenotypes. In order to preprocess the gene expression datasets, we relied on the procedure proposed in [47], applying the `recipe_zheng17` function from the Scanpy library [43].

Since our study aims to unravel the GRN between key regulators of AD in the SH-SY5Y cell-line model, we focused on a subset of genes of interest. First, we selected AD-related gene sets reported in collections C2 (curated), C3 (regulatory), C5 (ontology) and C7 (immunologic) of the MSIGDB database [34], using keywords detailed in Table 1. Then, we focused only on genes encoding

[4] `https://www.ebi.ac.uk/gxa/home`

Transcription Factors (TFs), i.e., proteins that can control the expression of other genes. In practice we used the lists of TFs provided in the AnimalTFDB database [32], and we focused on TFs that are preserved between mouse and human, i.e., for which homologs exist in both species (a one-to-one mapping between both species was ensured applying an identity-based reciprocal best Blast hit [41]). Finally, we removed genes that were not expressed in all samples from each of the four datasets presented in the previous paragraph. This gene selection procedure resulted in a final list of 386 genes of interest.

Table 1: Gene sets collections used to select genes of interest

Collect.	**keywords**
C2, C3, C7, C8	neuron, alzheimer, inflammation, neuroinflammation, wp tyrobp causal network in microglia, immune response, wp microglia pathogen phagocytosis pathway
C5	gobp (negative and positive) regulation of microglial cell activation, alzheimer, neuroinflammation, cytokine production involved in inflammatory response, cytokine production involved in immune response, gobp microglial cell activation involved in immune response, gobp microglial cell mediated cytotoxicity
Excluded	fetal, embryonic

3.2 CORAL domain adaptation

A fundamental assumption in ML is that the training and test sets should belong to the same domain. Nevertheless, gene expression matrices from different sources (e.g., different cell types or species) are likely to follow different distributions and thus belong to different domains. It is therefore unlikely that a GRN learned on a given data set using ML could be directly used to model a different data set. Furthermore, the inference of a GRN on a large set of AD-related gene expression matrices from disparate domains is likely to be unsuccessful, since the model would be more inclined to reflect inter-domain differences rather than the intrinsic variations between AD and control conditions. This problem has been reported in analogous scenarios, under the term "batch effect" [38]. In this study, we address this issue by adapting the gene expression between disparate datasets, thereby improving the quality of the inferred GRN. In practice, we applied the CORrelation ALignment (CORAL) algorithm [35] to align the genes expressions from a source domain with those from a target domain before inferring a GRN. CORAL is a simple and effective unsupervised domain adaptation method. It aims at applying a linear mapping to align the covariances of source and target distributions. Here, the covariance matrix represents the degree to which each pair of gene expressions is linearly related. In essence, the CORAL algorithm entails a linear transformation of the source gene expression, such

that the linear relationship between pairs of genes in terms of gene expression is consistent with that observed in the target data.

More formally, let $\mathbf{X} \in \mathbb{R}^{D \times N}$ denote a gene expression matrix, where D is the number of libraries, and $N = |\Gamma|$ is the number of genes. Let $\mathbf{X}_S \in \mathbb{R}^{D_S \times N}$ and $\mathbf{X}_T \in \mathbb{R}^{D_T \times N}$ be respectively the source and target datasets, with possibly different numbers of libraries, but described by the same set of genes Γ. Let $\mu_S \in \mathbb{R}^N$ and $\mu_T \in \mathbb{R}^N$ be the column-wise mean gene expression vectors and let $\mathbf{C}_S \in \mathbb{R}^{N \times N}$ and $\mathbf{C}_T \in \mathbb{R}^{N \times N}$ denote the gene expression covariance matrices $\mathbf{C}_S = Cov(\mathbf{X}_S)$ and $\mathbf{C}_T = Cov(\mathbf{X}_T)$. Let us assume that column vectors are Z-scored to exhibit a zero mean (i.e., $\mu_S = \mu_T = \mathbf{0}$) and unitary variance (i.e., $\sigma_S^2 = Diag(\mathbf{C}_S) = \mathbf{1}$ and $\sigma_T^2 = Diag(\mathbf{C}_T) = \mathbf{1}$ where $Diag(\mathbf{C}) \in \mathbb{R}^N$ denotes the diagonal vector of matrix $\mathbf{C}$). CORAL aims at minimizing $\|\mathbf{C}_{\hat{S}} - \mathbf{C}_T\|_F^2$, the Frobenius distance between $\mathbf{C}_T$ and $\mathbf{C}_{\hat{S}}$, where $\mathbf{C}_{\hat{S}} = \mathbf{A}^\top \mathbf{C}_S \mathbf{A}$ denotes the covariance matrix of $\mathbf{X}_{\hat{S}} = \mathbf{X}_S \mathbf{A}$, i.e., the source data $\mathbf{X}_S$ adapted through the linear mapping $\mathbf{A} \in \mathbb{R}^{N \times N}$, that can be expressed as $\mathbf{A} = \mathbf{C}_S^{-\frac{1}{2}} \mathbf{C}_T^{\frac{1}{2}}$ [35]. In practice, for the sake of stability, CORAL relies on regularized versions of the covariance matrices, i.e., $\mathbf{C}_S = Cov(\mathbf{X}_S) + \lambda \mathbf{I}$ and $\mathbf{C}_T = Cov(\mathbf{X}_T) + \lambda \mathbf{I}$, with $\lambda \in \mathbb{R}^+$ being the regularization parameter. The CORAL algorithm has shown to be very stable with respect to its regularization parameter for $0 \leq \lambda \leq 1$. Let CORAL be defined as a function $CORAL : \mathbb{R}^{D_S \times N} \times \mathbb{R}^{D_T \times N} \times [0, 1] \rightarrow \mathbb{R}^{D_S \times N}$ that receives as inputs $\mathbf{X}_S, \mathbf{X}_T$ and λ, and returns the adapted dataset $\mathbf{X}_{\hat{S}}$.

3.3 Semi-supervised label transfer

The diverse samples within a gene expression dataset are often characterised by a set of vectors of annotation, which describe the various aspects of the corresponding biological samples (e.g., species, tissue, cell type or pathology). More formally, let ${}^a\mathbf{Y} \in {}^a\mathcal{Y}^D$ be a vector with D elements from ${}^a\mathcal{Y}$, the set of possible values for the a-th characteristic (e.g., ${}^a\mathcal{Y} = \{"Male", "Female"\}$ for a denoting the sex). When a source dataset $\mathbf{X}_S$ is adapted into $\mathbf{X}_{\hat{S}}$ using CORAL, its corresponding label vectors should also be inferred in the target domain. To do so, we rely on the Label Spreading semi-supervised learning algorithm [48]. This method is based on the premise that samples with similar expression profiles should also exhibit comparable labels. For instance, it seems reasonable to posit that, following transfer by CORAL, a sample from an unsexed cell-line culture displaying a greater degree of similarity to the female patient samples, in the target domain, would be more appropriately labeled as "female."

More formally, let $\mathbf{X} = \begin{bmatrix} \mathbf{X}_T \\ \mathbf{X}_{\hat{S}} \end{bmatrix} \in \mathbb{R}^{(D_T + D_S) \times N}$ be the row augmented matrix built from matrices $\mathbf{X}_T$ and $\mathbf{X}_{\hat{S}}$. Moreover, let ${}^a\mathbf{Y} = [{}^a\mathbf{Y}_T \mid {}^a\mathbf{Y}_{\hat{S}}] \in ({}^a\mathcal{Y} \cup \{\varnothing\})^{D_T + D_S}$ be the labels of libraries in $\mathbf{X}$, for the a-th characteristic, such that its first D_T elements are equal to ${}^a\mathbf{Y}_T$, i.e. the labels of $\mathbf{X}_T$, and the remaining elements ${}^a\mathbf{Y}_{\hat{S}}$ correspond to the unlabeled samples from $\mathbf{X}_{\hat{S}}$ (here denoted with symbol $\varnothing$). Label Spreading is an iterative algorithm that builds a weighted fully connected graph where nodes denote samples (both labeled and unlabeled). The

weight $w_{i,j}$ between two nodes i and j representing samples with expression vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, is computed using the Radial Basis Function kernel $RBF : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+$ as $RBF(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \cdot \|\mathbf{x}_i - \mathbf{x}_j\|_2^2}$, with parameter $\gamma \in \mathbb{R}_*^+$. This function ensures high weights for libraries sharing similar expression profiles. Then, the weights between nodes are used to compute a transition probability $P(j \rightarrow i) = t_{i,j}$ between nodes by normalizing the weights: $t_{i,j} = \frac{w_{i,j}}{\sum_{k=1}^{D_T + D_S} w_{k,j}}$. Finally labels are spread to unlabeled samples using the transition probabilities until convergence. Let function $LabelSpreading : \mathbb{R}^{D_T \times N} \times \mathbb{R}^{D_S \times N} \times {}^a\mathcal{Y}_T^{D_T} \times \mathbb{R}_*^+ \rightarrow {}^a\mathcal{Y}_T^{D_S}$ receiving as input $\mathbf{X}_T, \mathbf{X}_S, {}^a\mathbf{Y}_T$ and γ, and returning the labeled vector ${}^a\mathbf{Y}_{\hat{S}}$, be the Label Spreading method.

3.4 GRN inference

GRN inference methods based on classification algorithms aim at training an ML model to predict the discretized level of expression of each gene (response variable), from the expressions of a set of regulatory genes (explanatory variables) [29]. More formally, let $g \in \Gamma$ denote a gene, and let $\mathbf{X}^{(g)} \in \mathbb{R}^D$ be its gene expression vector across D samples. Let $\Phi \subseteq \Gamma$ be a set of regulatory genes, such that $g \notin \Phi$, and let $\mathbf{X}^{(\Phi)} = \mathbb{R}^{D \times |\Phi|}$ denote their expression across D samples. Let $\mathbf{z}^{(g)} = Disc(\mathbf{X}^{(g)})$ denote the discretized vector of gene expressions of g, where $Disc : \mathbb{R}^D \rightarrow \mathcal{Z}^D$ is simply a discretization function mapping each continuous element of a vector into a set $\mathcal{Z} = \{0, 1, 2, \dots\}$, where $K = |\mathcal{Z}|$ denotes the number of possible classes. Let a classification model be a function $f^{(g)} : \mathbb{R}^{|\Phi|} \rightarrow \mathcal{Z}$, from a function space $\mathcal{F}$, that predicts, for each sample j, the discrete gene expression $\mathbf{z}_j^{(g)}$ of gene g from the regulators' gene expressions $\mathbf{X}_j^{(\Phi)}$. In this context, each regulator $\phi \in \Phi$ is considered as a predictive feature. A so-called training algorithm $R : \mathbb{R}^{D \times |\Phi|} \times \mathcal{Z}^D \rightarrow \mathcal{F}$ is used to find a model $f^{(g)}$ that minimizes some kind of empirical risk, that can be estimated here as the average classification error on the training dataset, i.e., $R\left(\mathbf{X}^{(\Phi)}, \mathbf{z}^{(g)}\right) = \underset{f^{(g)} \in \mathcal{F}}{argmin} \quad \frac{1}{D} \times \sum_j error\left(\mathbf{z}_j^{(g)}, f^{(g)}\left(\mathbf{X}_j^{(\Phi)}\right)\right)$.

3.5 Overall procedure

Given $\mathbf{X}_S$ and $\mathbf{X}_T$ a source and a target gene expression matrices, and given $\{{}^a\mathbf{Y}_T \ ; \ \forall a \in \{1, \dots, A\}\}$ a set of A annotation vectors describing each target library, Algorithm 1 describes the methodology used to transfer a gene expression dataset from a source to a target domain. First, the gene expression is mapped using the *CORAL* algorithm (line 2), and then all annotation labels are transferred using the *LabelSpreading* algorithm (line 3 to 5), finally the transferred gene expression matrix $\mathbf{X}_{\hat{S}}$, and the set of all transferred annotations $\{{}^a\mathbf{Y}_{\hat{S}} \ ; \ \forall a \in \{1, \dots, A\}\}$ are returned. The aforementioned schema is applicable to a simple scenario where only a source and a target data set exist. However, if there are multiple possible source data sets, the same

procedure can be applied iteratively to transfer all source data sets to a single source domain, before transferring it to the target domain. This is illustrated in Section 3.6. Following the transfer of the source dataset, it can be merged with the target dataset to create a unified dataset within the target domain, that can be used to infer the gene regulatory models. More formally, let $\mathbf{X} = \begin{bmatrix} \mathbf{X}_T \\ \mathbf{X}_{\hat{S}} \end{bmatrix} \in \mathbb{R}^{(D_T+D_S)\times N}$ be the row augmented matrix built from matrices $\mathbf{X}_T$ and $\mathbf{X}_{\hat{S}}$, and let $\{{}^a\mathbf{Y} = [{}^a\mathbf{Y}_T|\ {}^a\mathbf{Y}_{\hat{S}}];\ \forall a \in \{1,\ldots,A\}\}$ be the set of augmented vectors built from vectors ${}^a\mathbf{Y}_T$ and ${}^a\mathbf{Y}_{\hat{S}}$ for each kind of annotation $a \in \{1,\ldots,A\}$. Finally, Algorithm 2 is applied on the full dataset $\mathbf{X}$, in order to learn $\{f^{(g)};\ \forall g \in G\}$, the set of gene regulatory models for the set of genes G, with respect to the set of regulatory genes $\Phi \subseteq G$. A schematic representation of the overall pipeline is depicted in Figure 1.

Algorithm 1: Adapt Gene Expression Dataset

Input: $\mathbf{X}_T$, $\mathbf{X}_S$, $\{{}^a\mathbf{Y}_T\ ;\ \forall a \in \{1,\ldots,A\}\}$, λ, γ;
$\mathbf{X}_{\hat{S}} \leftarrow CORAL(\mathbf{X}_S, \mathbf{X}_T, \lambda)$;
for $a \leftarrow 1$ *to* A **do**
  ${}^a\mathbf{Y}_{\hat{S}} \leftarrow LabelSpreading(\mathbf{X}_T, \mathbf{X}_{\hat{S}}, {}^a\mathbf{Y}_T, \gamma)$;
end for
Return: $\mathbf{X}_{\hat{S}}$, $\{{}^a\mathbf{Y}_{\hat{S}}\ ;\ \forall a \in \{1,\ldots,A\}\}$;

Algorithm 2: Infer gene regulatory models

Input: $\mathbf{X}$, R, $Disc$, Φ, G;
$GRN \leftarrow \emptyset$;
for $g \in G$ **do**
  $\Phi' \leftarrow \Phi \setminus \{g\}$;
  $\mathbf{z}^{(g)} \leftarrow Disc(\mathbf{X}^{(g)})$;
  $f^{(g)} \leftarrow R(\mathbf{X}^{(\Phi')}, \mathbf{z}^{(g)})$;
  $GRN \leftarrow GRN \cup \{f^{(g)}\}$;
end for
Return: GRN;

3.6 Implementation and experimental protocol

In practice, the *CORAL* algorithm was implemented in Python, and the regularization parameter was set to $\lambda = 10^{-5}$. The *LabelSpreading* algorithm, from the scikit-learn library [4], was employed with the default parameters. In, order to infer GRNs from gene expression data, we relied on the GReNaDIne Python

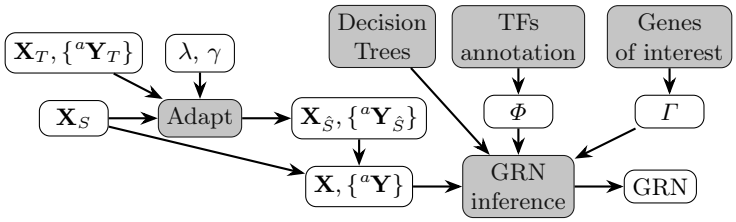


Fig. 1: Schema illustrating the TL-based approach used to align source and target data for downstream GRN inference.

package [30]. This library trains ML algorithms, from the scikit-learn library [4], to model the expression level of each gene as a function of the expression levels of other regulatory genes. In this work, we trained, for each gene, a DT to classify the gene level as *low* or *high*. The binarization of the gene expressions was ensured using the K-means algorithm, as suggested in [29]. In addition, we used GReNaDIne's clustering-based hyper-parameter tuning. This procedure clusters genes according to their gene expression level, and then, for each cluster, it performs a hyper-parameters grid-search exploration on the medoid gene, using a 3-fold cross-validation technique. Then, the selected hyper-parameters were used for all the genes in the same cluster, and hence only few genes were used for the hyper-parameter setting. Here, we explored a single hyper-parameter for the DTs, namely the `max_depth` $\in [1, 2, 3, 4, 5]$, to enforce simple trees, and prevent overfitting. The remaining parameters were set to their default values. The quality of each model was assessed, using a 3-fold cross-validation technique, based on standard classification metrics, namely the balanced accuracy, the Area Under the ROC curve (AUROC) and the F_1 score. We assessed the interpretability of the models, by computing their number of nodes. In order to compare the quality and interpretability of DTs w.r.t state-of-the-art black-box models, we also fit, for each gene, a Random Forest (RF) classifier, and a Support Vector Machine Classifier (SVM) as described in [29]. For the sake of reproducibility, our implementation is made available online[5].

In order to adapt the domains from the different datasets presented in Section 3.1, we decided to apply the iterative approach illustrated in Figure 2. First, the **Allen** dataset libraries from the *hippocampus* and the *white matter* tissues domain were transferred to the *neocortex* domain (step 1•A). Then, all libraries from **Allen** dataset, once fully transferred to the *neocortex* tissue domain, were transferred to the **Hisayama** neuron domain (step 1•B). Similarly, the libraries from the **PS2APP** dataset, corresponding to *astrocytes* and *glia* cell-types domains were transferred to the *neuron* domain (step 2•A). Then, all library from **PS2APP**, once transferred to the neuron cell type domain, were transferred to the **Hisayama** domain. Even if both datasets correspond to neural cell type, they have been obtained from different species, and thus this transformation aims at aligning the species differences (step 2•B). Finally, all libraries that were pre-

[5] `https://gitlab.com/bf2i/ad-coral`

viously transferred to the **Hisayama** domain were adapted to the **SH-SY5Y** cell line domain.

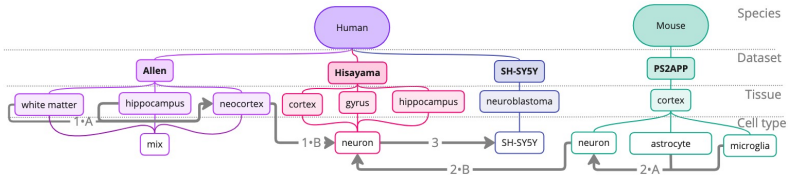


Fig. 2: Steps applied to transfer different datasets to the **SH-SY5Y** domain.

4 Results

4.1 Gene Expression Data Adaptation Qualitative Assessment

In order to assess the effectiveness of CORAL, we used the UMAP method [21] to represent the expression profiles of the different samples before and after each transfer step, as shown in Figure 3. Figure 3a shows that, once transferred towards the *neocortex* domain, **Allen** libraries from *hippocampus* and *white matter* lay in the same region as those from *neocortex*. Figure 3b shows an analogous behaviour for **CvsP2APP** libraries, transferred from *microglia* and *astrocyte* cell-type domains towards *neuron* cell-type domain. Similarly, as shown in Figures 3a and 3b, once mapped to the **Hisayama** neuron domain, all transferred libraries both from **Allen** and **CvsP2APP** datasets, lay close to those from the **Hisayama** dataset. Finally, Figure 3b shows that, all libraries laying on the **Hisayama** neuron domain, once transferred towards the **SH-SY5Y** one, lay next to target libraries, suggesting the effectiveness of CORAL despite the small size of the target dataset.

Figure 3d represents all the libraries once mapped to the **SH-SY5Y** domain, altogether with their corresponding AD conditions inferred via Label Spreading. Interestingly, control samples tend to cluster together in the upper part of the figure, next to control libraries, lay samples with low AD-like profile, while in the rest distribution lay libraries with medium and high AD-like profiles.

4.2 GRN Inference assessment

As presented in Section 3.4, DT, RF and SVM models were trained to classify the expression level of all genes, as a function of the levels of expression of other genes. The distribution of average test scores, for DT, RF and SVM gene expression models, on the original **SH-SY5Y** dataset, and on the transferred dataset are depicted in Figure 4a. This figure shows that for all models, the incorporation of transferred data increased substantially the models' quality

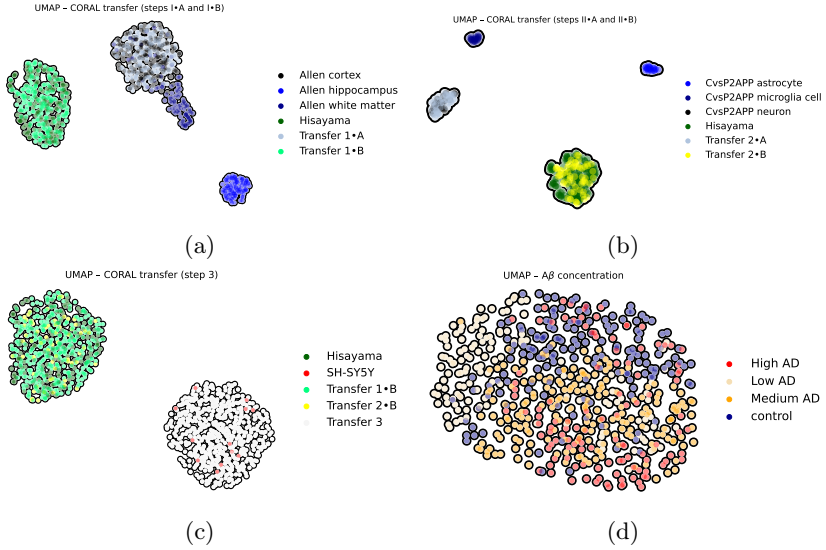
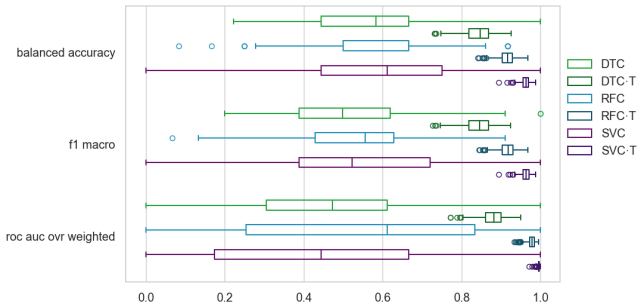


Fig. 3: UMAPs representing libraries expression profiles before and after transfer step 1 (3a), step 2 (3b) and step 3 (3c). UMAP representing libraries expression profiles transferred to the **SH-SY5Y** data domain, with the corresponding AD conditions inferred by the *LabelSpreading* method (3d).

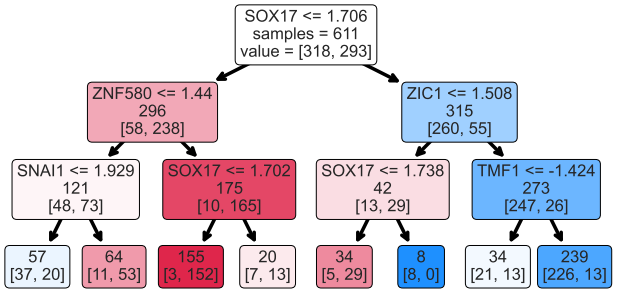
(higher than 0.3 for the different scores), and in this case both models exhibit suitable performances. Furthermore, as expected, sophisticated RF and SVM models have higher scores than DTs (around 0.1 for the different scores), but also exhibit complex structures, with 7783.29 nodes in average for RFs, 386 parameters for SVMs and only 10.94 nodes in average for simple DTs. In addition, the structures of DTs are particularly easy to read and interpret, as illustrated for instance in Figure 4b. Hence, in such a biomedical context, it is reasonable to opt for a limited loss of classification quality at the expense of a significant gain in terms of interpretability.

4.3 Biological interpretations

In order to better understand the GRN inferred, we assessed whether it tends to structure in communities that share common functional roles, as in [28]. To do so, we extracted communities from the GRN using the Clauset-Newman-Moore greedy modularity maximization technique [7], with default parameter settings. Twelve communities containing between 17 and 57 genes were extracted. Then, we evaluated the average level of collective over/under expression of genes from each community, in each condition using the `scanpy.tl.score_genes` function, as shown in Figure 5a. According to this Figure, community 5 is over-expressed in *High AD* and under-expressed in *Low AD*, while communities 8 and 11 are



(a) Average test distributions



(b) DT model for gene *RelA*

Fig. 4: Average test score distributions for RFs, SVMs and DTs trained on data including or not transferred instances (4a). Example of DT model for gene *RelA*, "high" and "low" expression classes are depicted in red and blue (4b).

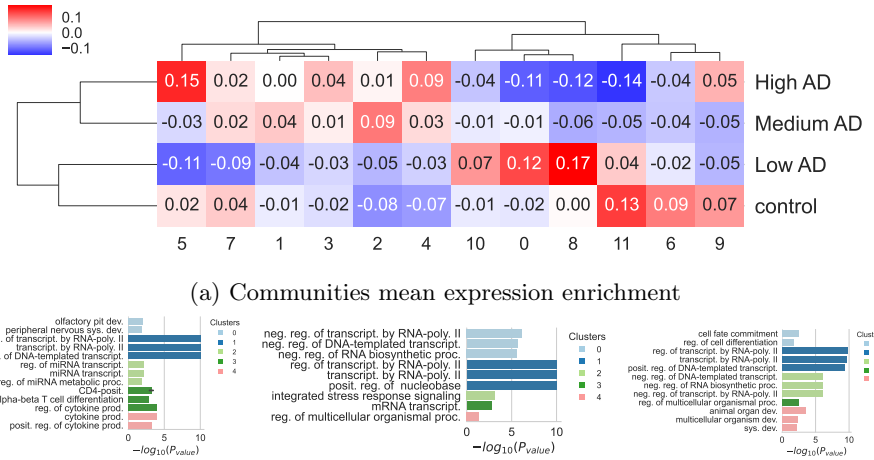
under-expressed in *High AD* and over-expressed in *Low AD* and *Control* conditions respectively. Regarding functional enrichment, all communities are involved in regulation of transcription (e.g., regulation of transcription by RNA-polymerase II), this was expected since only transcription factors were selected in this work. But these communities also exhibit distinctive enrichment profiles: Community 5 is involved in inflammation (e.g., cytokine production) and immune response (e.g., CD4 positive), which is coherent with an AD-context of general neuroinflammation (Figure 5b); Community 8 is involved in stress response, which could be associated to an early step in AD development (Figure 5c); Finally, community 11 is involved in development and differentiation (e.g., cell fate commitment, animal organ development), which is coherent with the maintenance of a cell and tissue differentiation state in the control condition (Figure 5d). Finally, some important regulators known to be involved in AD are present in community 5. Indeed, *RelA* is directly linked to AD, since it regulates *BACE1*, an enzyme involved in the production of $A\beta$ oligomers [6]. In addition, *RelA* is an important component of the NF-κB pathway, which has been shown to be widely involved in inflammatory responses in general, and to play a central role in AD [36]. Interestingly, the interpretable DT modeling *RelA* expression, shown in Figure 4b, points to other important regulators, such as *Sox17*. This gene is known to be involved in so-called Wnt/β-catenin signaling pathway [27], that revealed to be involved in a complex cross-talk with NF-κB [19]. Moreover, it has been shown that inactivating *Sox17* leads to a brain vascular perturbation and disruption of the blood–brain barrier [8], which has been suggested to play a role in cognitive impairment and AD pathogenesis [13]. Interestingly, Figure 4b shows that *Sox17* under-expression is associated to *RelA* over-expression, which suggests that AD is characterized by an antinomical pro-inflammatory cross-talk between NF-κB and Wnt/β-catenin pathways. This hypothesis could be assessed by downregulating *Sox17* by RNA interference in a cell model of Alzheimer's disease and evaluating the expression level of *RelA*.

All the results, have been made available online[6].

5 Conclusion

In this work we have proposed a method to infer interpretable GRNs from RNA-seq datasets with few samples. To do so, our method relies on CORAL, a linear feature-based domain adaptation method, and a semi-supervised label spreading algorithm, in order to transfer data from a source to a target RNA-seq domain, and then, train interpretable DTs on the transferred data. Using this technique, we built a first interpretable GRN of key regulators of AD in the SH-SY5Y cell-line model. We assessed the data transfer qualitatively using UMAP representations, and we evaluated the quality of the models using well established measures for classification, and compared their performance to state-of-the-art complex models, including RFs and SVMs. This comparison revealed that TL improves the quality of DTs, RFs and SVMs. Moreover, despite an expected loss

[6] `https://gitlab.com/bf2i/ad-coral`.



(a) Communities mean expression enrichment

(b) Community 5

(c) Community 8

(d) Community 11

Fig. 5: Collective over/under expression of each community's genes in each condition (5a). Enriched biological functions for communities 5, 8 and 11 (5c,5b,5d).

of quality w.r.t. complex models, DTs maintain decent scores, and allow to easily interpret the results. Further explorations based on communities extraction and enrichment analysis, revealed that the GRN involved in AD-like phenotype of SH-SY5Y cell lines exhibit inflammatory and immune functions. In addition, our method has enabled us to identify important, well-known regulatory pathways involved in AD and to highlight promising new regulatory genes that deserve to be investigated experimentally in future work. The method proposed here is not limited to AD, and the suitable results obtained here suggest that it could be applied to other small datasets. Nevertheless, as stated in [46], TL is not guaranteed to be effective in every situation: Major differences between the source and target domains, poor quality (e.g., noise, outliers) in the source or target datasets, as well as deficiencies in the TL methods, can lead to negative transfer. To this end, important future work includes performing an in-depth sensitivity analysis on different datasets to assess the impact of different hyperparameters on the final results, as well as exploring the limitations imposed by the quality and quantity of available data, especially for small datasets. Furthermore, another promising future research direction is the exploration of more advanced domain adaptation techniques, such as adversarial domain adaptation, to potentially improve the alignment between source and target domains.

Acknowledgement

This work was supported by the BQR INSA Lyon 2023 NeurInfO, the French National Institute for Agriculture, Food, and Environment (INRAE), and the

National Institute for Applied Sciences (INSA Lyon). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Aibar, S., González-Blas, C.B., Moerman, et al.: Scenic: single-cell regulatory network inference and clustering. Nature methods **14**(11), 1083–1086 (2017)
2. Ali, M., Huarte, O.U., Heurtaux, et al.: Single-cell transcriptional profiling and gene regulatory network modeling in tg2576 mice reveal gender-dependent molecular features preceding alzheimer-like pathologies. Molecular Neurobiology **61**(2), 541–566 (2024)
3. Bhadra, S., Kelkar, V.A., Brooks, F.J., Anastasio, M.A.: On hallucinations in tomographic image reconstruction. IEEE transactions on medical imaging **40**(11), 3249–3260 (2021)
4. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., et al.: API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning. pp. 108–122 (2013)
5. Castellani, R.J., Rolston, R.K., Smith, M.A.: Alzheimer disease. Disease-a-month: DM **56**(9), 484 (2010)
6. Chen, C.H., Zhou, W., Liu, S., Deng, Y., et al.: Increased nf-κb signalling up-regulates bace1 expression and its therapeutic potential in alzheimer's disease. International Journal of Neuropsychopharmacology **15**(1), 77–90 (2012)
7. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Physical review E **70**(6), 066111 (2004)
8. Corada, M., Orsenigo, F., Bhat, G.P., et al.: Fine-tuning of sox17 and canonical wnt coordinates the permeability properties of the blood-brain barrier. Circulation research **124**(4), 511–525 (2019)
9. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., Wang, B.: scgpt: toward building a foundation model for single-cell multi-omics using generative ai. Nature Methods pp. 1–11 (2024)
10. Feng, G., Qin, X., Zhang, J., Huang, W., et al.: Cellpolaris: Decoding cell fate through generalization transfer learning of gene regulatory networks. bioRxiv pp. 2023–09 (2023)
11. Hokama, M., Oka, S., Leon, J., et al.: Altered expression of diabetes-related genes in alzheimer's disease brains: the hisayama study. Cerebral Cortex (New York, N.Y. : 1991) **24**(9), 2476–2488 (2014). https://doi.org/10.1093/cercor/bht101
12. Huynh-Thu, V.A., Sanguinetti, G.: Gene regulatory network inference: an introductory survey. Gene regulatory networks: Methods and protocols pp. 1–23 (2019)
13. Jefferies, W.A., Price, K.A., Biron, K.E., Fenninger, F., Pfeifer, C.G., Dickstein, D.L.: Adjusting the compass: new insights into the role of angiogenesis in alzheimer's disease. Alzheimer's research & therapy **5**, 1–9 (2013)
14. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023)
15. Khullar, S., Wang, D.: Predicting brain-regional gene regulatory networks from multi-omics for alzheimer's disease phenotypes and covid-19 severity. Human Molecular Genetics **32**(11), 1797–1813 (2023)

16. Lane, C.A., Hardy, J., Schott, J.M.: Alzheimer's disease. European Journal of Neurology **25**(1), 59–70 (2018). https://doi.org/10.1111/ene.13439, https://onlinelibrary.wiley.com/doi/full/10.1111/ene.13439
17. Levine, M., Davidson, E.H.: Gene regulatory networks for development. Proceedings of the National Academy of Sciences **102**(14), 4936–4942 (2005)
18. Littman, R., Cheng, M., Wang, N., Peng, C., Yang, X.: Scing: Inference of robust, interpretable gene regulatory networks from single cell and spatial transcriptomics. Iscience **26**(7) (2023)
19. Ma, B., Hottiger, M.O.: Crosstalk between wnt/β-catenin and nf-κb signaling pathway during inflammation. Frontiers in immunology **7**, 221254 (2016)
20. Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Kellis, M., Collins, J.J., et al.: Wisdom of crowds for robust gene network inference. Nature methods **9**(8), 796–804 (2012)
21. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint (2018)
22. de Medeiros, L.M., De Bastiani, M.A., Rico, E.P., Schonhofen, P., et al.: Cholinergic differentiation of human neuroblastoma SH-SY5Y cell line and its potential use as an in vitro model for alzheimer's disease studies. Mol. Neurobiol. **56**(11), 7355–7367 (2019)
23. Mignone, P., Pio, G., D'Elia, D., Ceci, M.: Exploiting transfer learning for the reconstruction of the human gene regulatory network. Bioinformatics **36**(5), 1553–1561 (2020)
24. Miller, J.A., et al.: Neuropathological and transcriptomic characteristics of the aged brain. eLife **6**, e31126 (2017). https://doi.org/10.7554/eLife.31126
25. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
26. Molnar, C.: Interpretable Machine Learning. 2 edn. (2022), https://christophm.github.io/interpretable-ml-book
27. Mukherjee, S., Chaturvedi, P., Rankin, S.A., Fish, M.B., et al.: Sox17 and β-catenin co-occupy wnt-responsive enhancers to govern the endoderm gene regulatory network. Elife **9**, e58029 (2020)
28. Peignier, S., Calevro, F.: Gene self-expressive networks as a generalization-aware tool to model gene regulatory networks. Biomolecules **13**(3), 526 (2023)
29. Peignier, S., Schmitt, P., Calevro, F.: Data-driven gene regulatory networks inference based on classification algorithms. International Journal on Artificial Intelligence Tools **30**(04), 2150022 (2021)
30. Schmitt, P., Sorin, B., Frouté, T., Parisot, N., et al.: Grenadine: a data-driven python library to infer gene regulatory networks from gene expression data. Genes **14**(2), 269 (2023)
31. Sebastian, S., Roy, S., Kalita, J.: A generic parallel framework for inferring large-scale gene regulatory networks from expression profiles: application to alzheimer's disease network. Briefings in Bioinformatics **24**(1), bbac482 (2023)
32. Shen, W.K., Chen, S.Y., Gan, Z.Q., Zhang, Y.Z., et al.: Animaltfdb 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. Nucleic Acids Research **51**, D39–D45 (Jan 2023)
33. Srinivasan, K., Friedman, B.A., Larson, J.L., Lauffer, B.E., et al.: Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses. Nature communications **7**(1), 11295 (2016)
34. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., et al.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expres-

sion profiles. Proceedings of the National Academy of Sciences **102**(43), 15545–15550 (2005). https://doi.org/10.1073/pnas.0506580102, https://www.pnas.org/doi/abs/10.1073/pnas.0506580102
35. Sun, B., Feng, J., Saenko, K.: Correlation Alignment for Unsupervised Domain Adaptation. arXiv e-prints arXiv:1612.01939 (Dec 2016). https://doi.org/10.48550/arXiv.1612.01939
36. Sun, E., Motolani, A., Campos, L., Lu, T.: The pivotal role of nf-kb in the pathogenesis and therapeutics of alzheimer's disease. International journal of molecular sciences **23**(16), 8972 (2022)
37. Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., et al.: Transfer learning enables predictions in network biology. Nature **618**(7965), 616–624 (2023)
38. Tung, P.Y., Blischak, J.D., Hsiao, C.J., Knowles, D.A., et al.: Batch effects and the effective design of single-cell gene expression studies. Scientific reports **7**(1), 39921 (2017)
39. Uhrig, M., Ittrich, C., Wiedmann, V., Knyazev, Y., et al.: New alzheimer amyloid β responsive genes identified in human neuroblastoma cells by hierarchical clustering. PLoS One **4**(8), e6779 (2009)
40. Wang, X.D., Liu, S., Lu, H., Guan, Y., et al.: Analysis of shared genetic regulatory networks for alzheimer's disease and epilepsy. BioMed Research International (2021)
41. Ward, N., Moreno-Hagelsieb, G.: Quickly finding orthologs as reciprocal best hits with blat, last, and ublast: how much do we miss? PLoS One **9**(7), e101850 (2014). https://doi.org/10.1371/journal.pone.0101850
42. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big data **3**, 1–40 (2016)
43. Wolf, F.A., Angerer, P., Theis, F.J.: Scanpy: large-scale single-cell gene expression data analysis. Genome biology **19**, 1–5 (2018)
44. Yang, Y., Bazhin, A.V., Werner, J., Karakhanova, S.: Reactive oxygen species in the immune system. International reviews of immunology **32**(3), 249–270 (2013)
45. Yiannopoulou, K.G., Papageorgiou, S.G.: Current and future treatments in alzheimer disease: an update. Journal of central nervous system disease **12** (2020)
46. Zhang, W., Deng, L., Zhang, L., Wu, D.: A survey on negative transfer. IEEE/CAA Journal of Automatica Sinica **10**(2), 305–329 (2022)
47. Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., et al.: Massively parallel digital transcriptional profiling of single cells. Nature communications **8**(1), 14049 (2017)
48. Zhou, D., Bousquet, O., Lal, T., Weston, J., Schölkopf, B.: Learning with local and global consistency. Advances in neural information processing systems **16** (2003)