frontiers | Frontiers in Cellular Neuroscience

TYPE Original Research
PUBLISHED 23 May 2024
DOI 10.3389/fncel.2024.1369242

OPEN ACCESS

EDITED BY
Ning Huang,
Chongqing Medical University, China

REVIEWED BY
Jiahe Tan,
First Affiliated Hospital of Chongqing Medical University, China
Austin W. T. Chiang,
Augusta University, United States

*CORRESPONDENCE
Fuhai Li
✉ fuhai.li@wustl.edu

†These authors share first authorship

RECEIVED 11 January 2024
ACCEPTED 30 April 2024
PUBLISHED 23 May 2024

CITATION
Feng J, Song H, Province M, Li G, Payne PRO, Chen Y and Li F (2024) PathFinder: a novel graph transformer model to infer multi-cell intra- and inter-cellular signaling pathways and communications.
Front. Cell. Neurosci. 18:1369242.
doi: 10.3389/fncel.2024.1369242

PathFinder: a novel graph transformer model to infer multi-cell intra- and inter-cellular signaling pathways and communications

Jiarui Feng[1,2†], Haoran Song[1,2†], Michael Province[3], Guangfu Li[4,5,6], Philip R. O. Payne[1], Yixin Chen[2] and Fuhai Li[1,7*]

[1]Institute for Informatics (I2), Washington University School of Medicine, Washington University in St. Louis, St. Louis, MO, United States, [2]Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO, United States, [3]Division of Statistical Genomics, Department of Genetics, Washington University in St. Louis, St. Louis, MO, United States, [4]Department of Surgery, University of Missouri-Columbia, Columbia, MO, United States, [5]Department of Molecular Microbiology and Immunology, University of Missouri-Columbia, Columbia, MO, United States, [6]NextGen Precision Health Institute, University of Missouri-Columbia, Columbia, MO, United States, [7]Department of Pediatrics, Washington University School of Medicine, Washington University in St. Louis, St. Louis, MO, United States

Recently, large-scale scRNA-seq datasets have been generated to understand the complex signaling mechanisms within the microenvironment of Alzheimer's Disease (AD), which are critical for identifying novel therapeutic targets and precision medicine. However, the background signaling networks are highly complex and interactive. It remains challenging to infer the core intra- and inter-multi-cell signaling communication networks using scRNA-seq data. In this study, we introduced a novel graph transformer model, PathFinder, to infer multi-cell intra- and inter-cellular signaling pathways and communications among multi-cell types. Compared with existing models, the novel and unique design of PathFinder is based on the divide-and-conquer strategy. This model divides complex signaling networks into signaling paths, which are then scored and ranked using a novel graph transformer architecture to infer intra- and inter-cell signaling communications. We evaluated the performance of PathFinder using two scRNA-seq data cohorts. The first cohort is an APOE4 genotype-specific AD, and the second is a human cirrhosis cohort. The evaluation confirms the promising potential of using PathFinder as a general signaling network inference model.

KEYWORDS

Alzheimer's disease, signaling pathways, cell cell signaling communications, microenvironment, graph neural network

Introduction

Single-cell RNA sequencing data (scRNA-seq) technologies have become popular in recent years because of their ability to profile gene expression and analyze cell composition in the single cell resolution (Kolodziejczyk et al., 2015; Tanay and Regev, 2017; Hwang et al., 2018). On the one hand, by profiling and annotating scRNA-seq data, researchers can analyze

differentially expressed genes in each cell population and sub-population to understand which gene is altered in certain conditions. On the other hand, scRNA-seq data also show great potential in discovering intra- and inter-cellular communication. However, there are only limited methods for discovering active signaling pathways or intra-cellular communication using scRNA-seq data. The existing models are mainly based on correlation, regression, and Bayesian analysis (Saint-Antoine and Singh, 2019), and the direct interaction signaling cascades were usually ignored in those methods because only a small set of genes exhibit gene expression changes between different conditions (Feng et al., 2020). For example, CellPhoneDB (Efremova et al., 2020) can model the interactions between ligands from one cell type and receptors from another cell type. However, it cannot model the downstream signaling. CCCExplorer (Choi et al., 2015) can discover both the ligand–receptor interaction and downstream the signaling network by modeling differentially expressed genes. NicheNet (Browaeys et al., 2020) takes a further step by integrating various interaction databases and training a predictive model to assess the interaction potential between the ligand and downstream targets. However, it only applies a statistical model, which cannot generate a clear communication path. CytoTalk (Hu et al., 2021) applies the Steiner tree to discover the de-novo signal transduction network from gene co-expression. However, the discovered signaling is based on co-expression, and the physical interaction cascade is still unknown.

In the past few years, graph neural networks (GNNs) have become famous due to their great performance in node and graph representation as wells as in classification tasks. For instance, GraphSAGE (Hamilton et al., 2017) proposed the first general framework for learning the node representation inductively. GAT (Veličković et al., 2017) incorporates the attention mechanism into GNNs to actively learn how to aggregate all the information in graphs. The DGCNN (Zhang et al., 2018) model proposes sortPooling to efficiently sort nodes and learn graph features for graph classification. GIN (Xu et al., 2018) connects message-passing GNNs with the 1-dimensional Wifelier-Lehman test (1-WL test) on learning graph structure and proposes a new GNN algorithm that is equally powerful as the 1-WL test. More recently, researchers have tried to generalize the transformer architecture (Vaswani et al., 2017) into graph learning fields as it already shows superior power in learning both text and image data. Many studies (Cai and Lam, 2020; Hu et al., 2020; Rong et al., 2020; Zhang et al., 2020; Yang et al., 2021; Ying et al., 2021) have shown great potential in applying the transformer model to the graph data. They either nest GNN architectures in the transformer layer, design specific attention mechanisms, or design novel encoding mechanisms to incorporate the graph structure into the transformer model. However, using GNNs to discover the intra- and inter-cell communication network remains unknown as these networks are typically black-box models and it is hard to interpret their prediction results.

In this study, we present a novel framework called PathFinder to discover both intra- and inter-cell communication networks with a novel graph transformer-based neural network. Given the scRNA-seq expression data and the condition (control/test), PathFinder first samples a series of predefined paths through the prior gene–gene interaction database. Then, the PathFinder model takes the scRNA-seq expression data and the predefined path list as inputs to predict the condition of each cell. Through the training, the path important score will be learned to indicate the relative importance of each path in separating between the control and test conditions. To learn different types of communication, such as upregulated or downregulated networks, a novel regularization term is introduced. PathFinder will first generate a prior score for each path based on the expression level of genes in the path. Then, during the training, this regularization term will regularize the learned path scores to be close to the prior scores. After training, the path score will be sorted and the intra-communication network for each cell type will be generated by extracting the top K important paths. To generate the inter-cell communication network between the ligand cell and the receptor cell, the intra-cell communication network for the receptor cell will be collected, and the ligand list will be extracted from the differential expressed gene list in the ligand cell. Finally, the ligands are linked to the intra-cell network based on the ligand–receptor interaction database. The overall procedure of generating both intra- and inter-cell communication networks using PathFinder is shown in Figure 1. To thhe best of our knowledge, this is the first method to apply deep learning and graph transformers to discover signaling networks in scRNA-seq data. The advantages of PathFinder are listed below: (1) The model is designed based on a graph transformer, which has the great ability to learn both local and long-range signaling patterns from gene expression and large-scale networks. (2) It is capable of identifying and providing the full signaling network between cells via cellular ligands and receptors. (3) The proposed PathFinder is a general framework that allows users to input their own defined signaling paths or gene–gene interaction network database to identify important signaling based on their interests. Furthermore, (4) it can separate and generate different types of communication networks (Differential expressed/upregulated/downregulated), which allows more precise downstream analysis. We applied the PathFinder model on two scRNA-seq data cohorts: one is a mice cohort of AD and another is a human cohort of cirrhosis. The PathFinder not only achieves great prediction results but also generates intra- and inter-cell communication networks that align well with the latest knowledge on the mechanism of both two diseases.

Results

scRNA-seq data of Alzheimer's disease cohort on mice

To evaluate the proposed PathFinder method, scRNA-seq data on Alzheimer's disease are collected from the Gene Expression Omnibus (GEO) database with accession number GSE164507 (Wang et al., 2021). The raw data are processed using the Seurat R package (Hao et al., 2021), and the process procedure is conducted by following the previous study's procedure (Wang et al., 2021). Specifically, we select cell samples from two different conditions, denoted as TAFE4_tam and TAFE4_oil. TAFE4_tam refers to mice with the APOE4 gene knocked out from astrocyte cells, and TAFE4_oil refers to mice with the existence of APOE4. It is well known that APOE4 is one of the most significant genetic risk factors for late-onset AD. By analyzing the difference between the signaling pattern with and without APOE4, we can gain a deeper understanding of the effects of the APOE4 gene on brain cells.

10.3389/fncel.2024.1369242

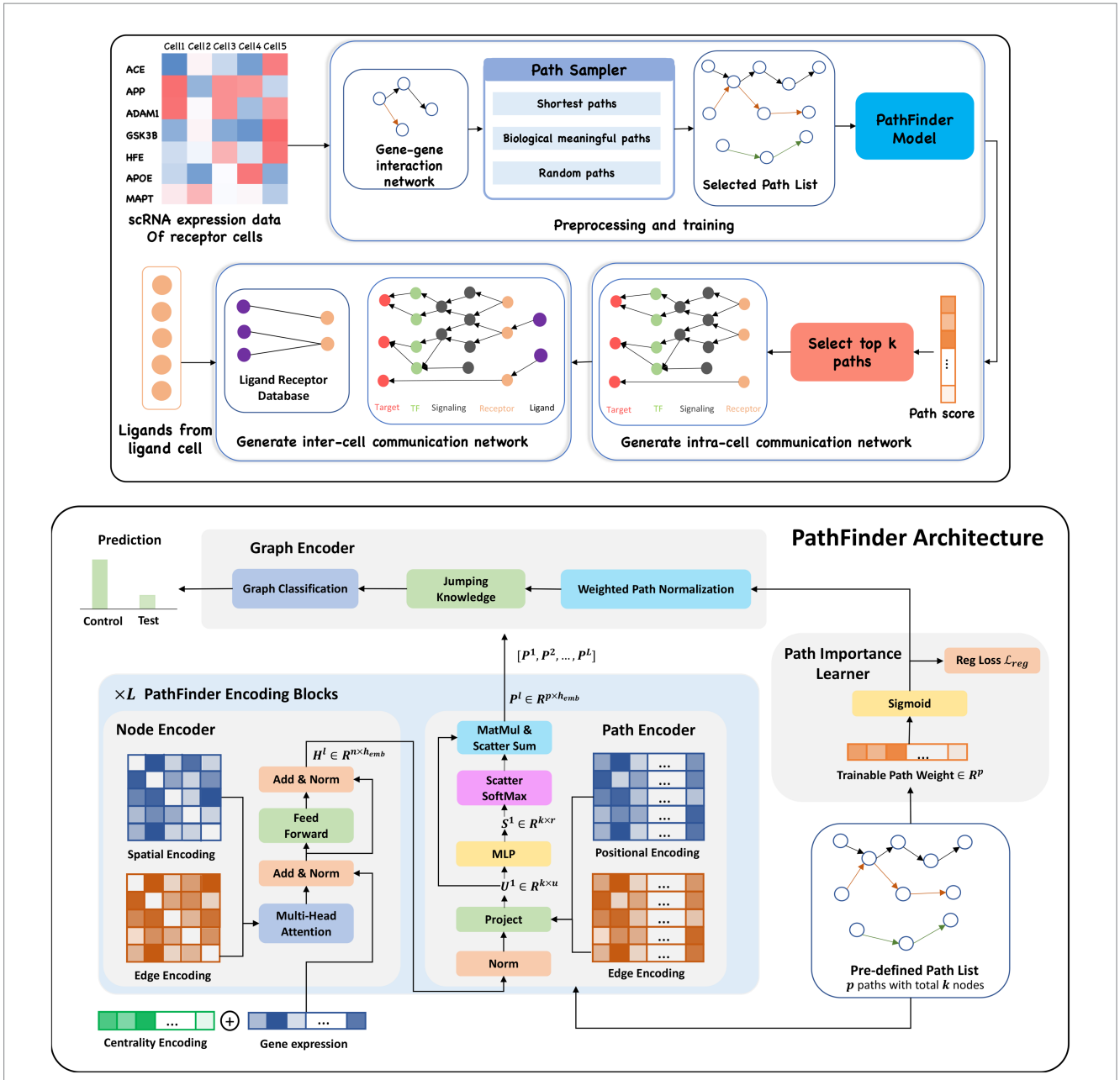

FIGURE 1
(Upper) Overview of the PathFinder method to discover both intra- and inter-cell communication networks. The input scRNA expression data with both samples from the control condition and the test condition are used to construct the gene–gene interaction network based on our large database. Then, the path sampler is used to generate all pre-defined path from the interaction network. Then, the PathFinder model is trained to separate the cells from two different conditions. After the training, the learned path score can indicate the importance of each path. The top k paths are selected to generate the intra-cell communication network. Finally, the ligand–receptor database is used to link all picked ligands (like differential expressed ligands) from ligand cells to the receptors in the intra-cell communication network of receptor cells to construct the inter-cell communication network. **(Lower)** Model architecture of PathFinder. The PathFinder model consists of three components: node encoder, path encoder, and graph encoder. The node encoder is a stack of L the transformer layer with special encoding to encode local graph structure information of each node. The path encoder take the output from each layer of node encoder to learn long-range path embedding for each pre-defined path. Finally, the graph encoder aggregate information from each path to generate graph embedding and make final prediction. In the graph encoder, the trainable path weight will be learned to assign each path an importance score, which can be used to generate intra-cell communication networks.

Concretely, the excitatory neuron (Ex), microglia (Mic), and astrocyte (Ast) of the TAFE4 group are collected from the dataset with a total number of samples of 13,604, 3,874, and 734, respectively. The detailed data distribution are provided in Supplementary Table S1. Then, the PathFinder method is applied to predict the condition of each cell (oil or tam) separately for each cell type and generate both intra- and inter-cell communication networks between these three cell types. The pre-defined path list includes all shortest distance paths starting from receptors and all possible paths from the receptor to the target gene. For the shortest distance paths, we only select paths with a minimum length of 3 (except all receptor direct regularizations, which have a length of 2) and a maximum length of 10. We compute the prior score of each path

based on the average differential expression level of all genes in the path (more details in the Method section) for the path score regularization. To ensure the robustness of the analysis, we only selected the top 8,192 variable genes from the original dataset as input to the model, which resulted in a final count of 1,210 pre-selected paths. The detailed path selection procedure can be found in the Method section.

scRNA-seq data of cirrhosis cohort on humans

The scRNA-seq data of human cirrhosis is obtained from the GEO database under the accession number GSE136103, which includes non-parenchymal cells collected from healthy individuals and patients with cirrhosis. After processing, single-cell data were obtained from five healthy individuals (healthy1-5) and five patients with cirrhosis (cirrhotic1-5). Similarly, the raw data are processed using the Seurat R package (Hao et al., 2021). After the process, we select three important cell types: endothelial (Endo), macrophages (Mac), and T cells (Tcell). The total number of cells for each cell type is 6,197, 9,173, and 20,950, respectively. The detailed data distribution is provided in Supplementary Table S1. Similar to the AD cohort, we use PathFinder to predict the cell condition for each cell type. The pre-defined path list is selected in the same way as the AD dataset. For the cirrhosis cohort, we selected the top 12,000 variable genes from the original dataset as input to the model, which resulted in a final count of 1,549 pre-selected paths.

PathFinder can effectively separate cells from different conditions of AD by selecting differentially expressed signaling paths

To evaluate the performance of the PathFinder model, it is applied to excitatory neurons, astrocytes, and microglia cells from the AD cohort separately to predict the conditions of each cell (tam/oil), denoted as TAFE4_ex, TAFE4_mic, and TAFE4_ast, respectively. For each cell type, we repeat the training five times, each time randomly splitting the whole dataset into train, validation, and test subsets at a ratio of 0.7/0.1/0.2. We report the average performance and standard deviation on the test set over all five runs. The detailed experimental setting can be found in the Method section. The detailed results are shown in Table 1 and Figure 2A.

As can be seen, the PathFinder can successfully classify the majority of cells in the test dataset into the correct condition. This means that, after training, the model learned the most important difference between the two conditions from a huge gene expression profile. Such differences can be reflected in the important score of each path, as the final prediction is made based on the different predefined paths. Among all results, the standard deviation of the metrics for TAFE4_ast is much larger than the other two cell types. We speculate that this discrepancy is caused by the limited number of cell samples in the TAFE4_ast group, which makes the model easily overfit to the training data.

Then, we evaluate the learned path score from each group. For each cell group, we first average the learned path score from five repeated runs to get the final path score. We average the absolute fold-change level of all genes within each path to get an average differential expression level for each path. Then, we compare the top 200 selected paths from the results of the PathFinder model to the remaining paths. The results are shown in Figure 2B. We can see that, for all three different cell types, the selected top 200 paths from PathFinder have a much higher average differential expression level compared to the remaining paths. The results indicate that PathFinder is effective in ranking differential expressed paths through the training. This can be attributed to two objective functions used in PathFinder. First, by minimizing the classification loss, the model is forced to increase the score for paths that are useful for separating two different conditions. It is intuitive that paths with higher average differential expression levels are more helpful for the prediction. Second, by minimizing the regularization loss, the model tends to give a high score for paths with high prior weight, and the prior weight is positively related to the average differential expression level.

Then, we evaluate the robustness and stability of the PathFinder. Concretely, we want the final path score distribution (ranking) learned from PathFinder to be stable and robust even if we slightly alter the training data. Since we randomly split the whole dataset for each repeated run, we can directly compare the learned score for each run to achieve our goal. Therefore, we plot the learned score for all paths, and all runs with paths are sorted by the average score. The results are shown in Figure 2C. For all three cell types, the learned scores are very stable across different runs, as paths with higher ranks always have higher scores. This means that, even if we slightly alter the training dataset, the PathFinder model can still output almost the same top k paths. The results successfully demonstrate the robustness of the PathFinder model for extracting important paths and constructing intra-communication networks.

Finally, we further evaluate the effectiveness of PathFinder on intra-cell signaling networks using the human cirrhosis cohort. Specifically, we run PathFinder on endothelial, macrophages, and T cells. The procedure is the same as the AD cohort. The average evaluation metric on the test set can be found in Supplementary Table S2 and the comparison of the average differential expression level of paths can be found in Supplementary Figures S1A,B.

Core intra-cell signaling networks associated with the APOE4 genotype

In this section, we evaluate the intra-cell communication networks discovered by the PathFinder model. Particularly, we want to know

TABLE 1 Evaluation results of the PathFinder model.

	Accuracy	Recall	Precision	Specificity	F1	AUC
TAFE4_ex	0.67 ± 0.01	0.71 ± 0.04	0.66 ± 0.02	0.64 ± 0.05	0.68 ± 0.01	0.73 ± 0.01
TAFE4_mic	0.67 ± 0.01	0.76 ± 0.03	0.65 ± 0.02	0.58 ± 0.04	0.70 ± 0.01	0.71 ± 0.01
TAFE4_ast	0.62 ± 0.04	0.75 ± 0.15	0.65 ± 0.03	0.44 ± 0.14	0.69 ± 0.06	0.65 ± 0.04

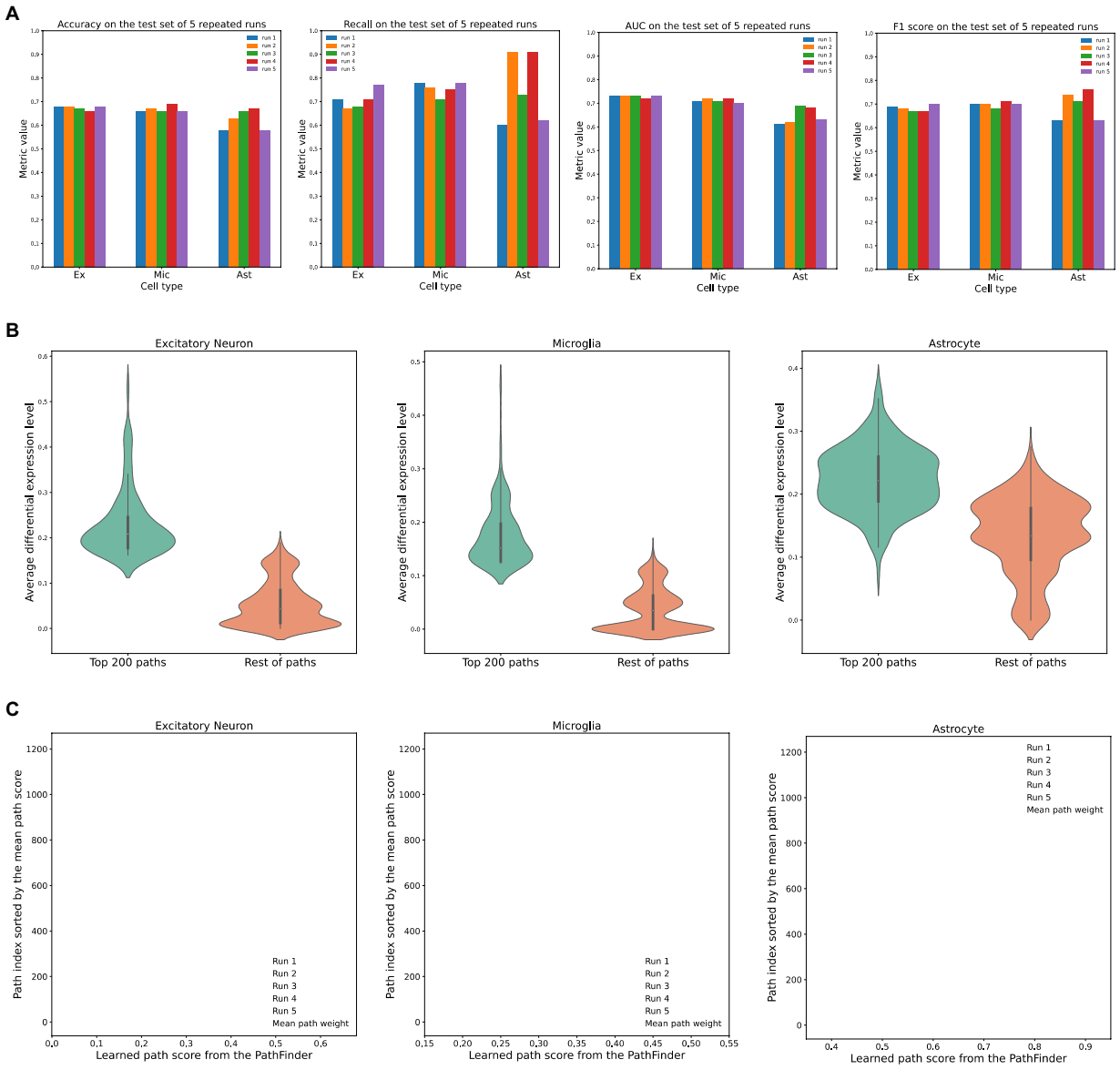

FIGURE 2
Evaluation of the PathFinder model on the AD cohort. **(A)** The detailed evaluation metrics on test dataset from all runs. **(B)** The comparison of the average differential expression level of top paths sorted by PathFinder during the training. The top 200 paths have higher differential expression level than others for all three cell types. **(C)** The learned path scores of PathFinder on different runs. All paths are ranked by the average score across all runs.

whether the discovered networks can reveal the recent discovery of APOE4-driven AD or even indicate new findings. First, for all three cell types, the final networks are generated by first averaging the path score learned from five repeated runs and then ranking and selecting the top 300 paths from all paths to form the final networks. The generated networks for all three cell types are shown in Figure 3. Then, we perform the enrichment analysis on all generated networks using KEGG signaling pathways and gene ontology (GO) terms. The enrichment results are shown in Figure 4A. Based on the results, we find several key factors that are important to the development of APOE4-driven AD.

Neuron inflammation

Numerous studies have shown that inflammation is highly activated and plays a key role in the progress of AD (Rogers et al., 1996; Akiyama et al., 2000; Halliday et al., 2000; Mathys et al., 2019). From the enrichment results, we can see that many inflammation-related pathways/GO terms are enriched across multiple cell types. For example, *cytokine-mediated signaling pathway, cellular response to cytokine stimulus,* and *inflammatory mediator regulation of TRP channels*. This result aligns with the findings of previous studies and further confirms that the existence of APOE4 in the astrocyte stimulates the inflammatory response. More specifically, several genes related to neuron inflammation are identified by PathFinder across multiple cell types. STAT1 and STAT3 are identified as hub genes connected to multiple targets in both the network of neurons and microglia. It has been shown that STAT1 plays a key role in regulating inflammatory responses and cellular death (Hu et al., 2002; Butturini et al., 2018). Moreover, the differential expression analysis (Figure 4B) reveals that STAT1 is

10.3389/fncel.2024.1369242

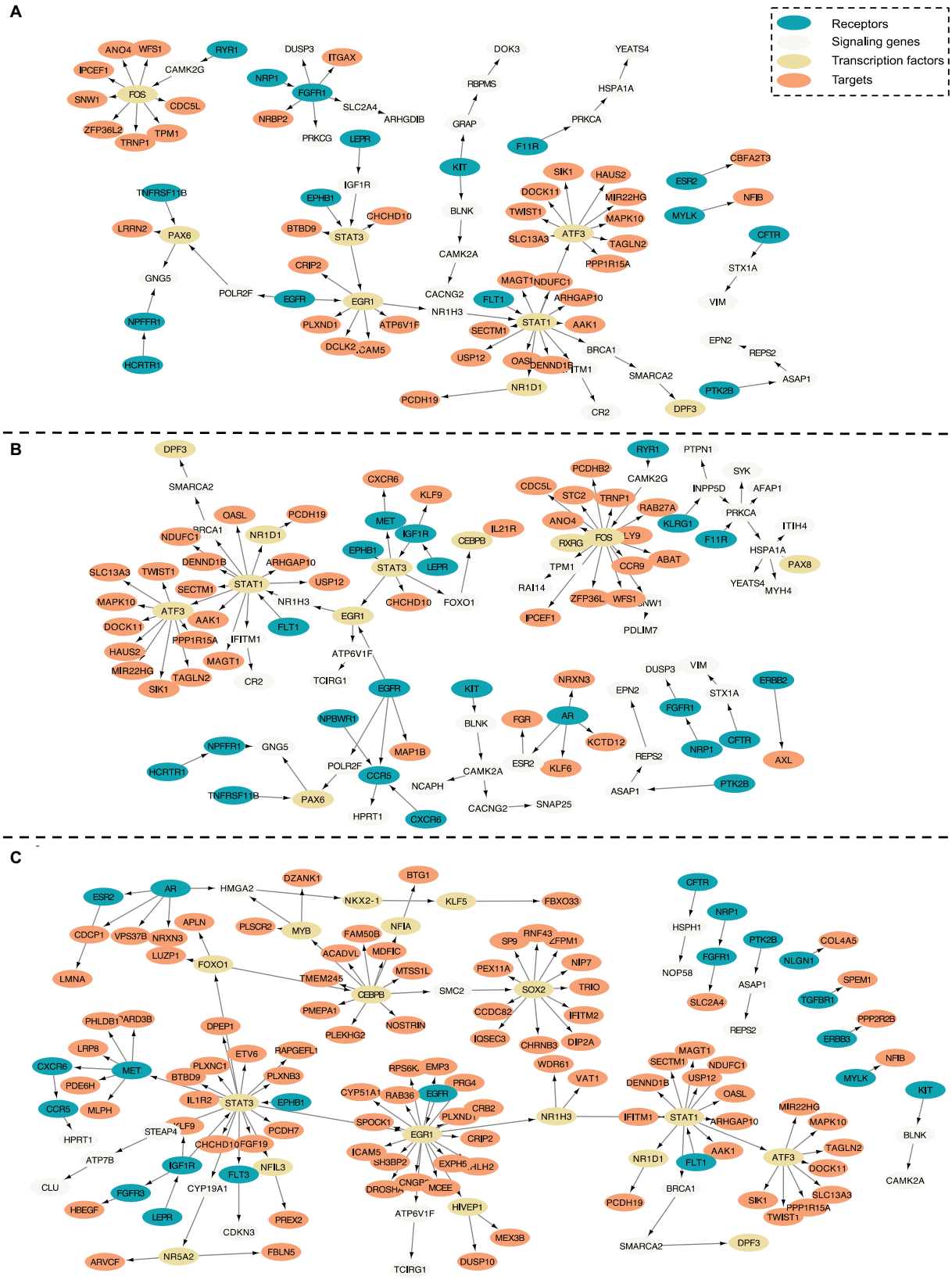

FIGURE 3
Intra-cell communication networks discovered by the PathFinder model for the AD cohort. **(A)** Excitatory neurons; **(B)** Microglia; **(C)** Astrocyte.

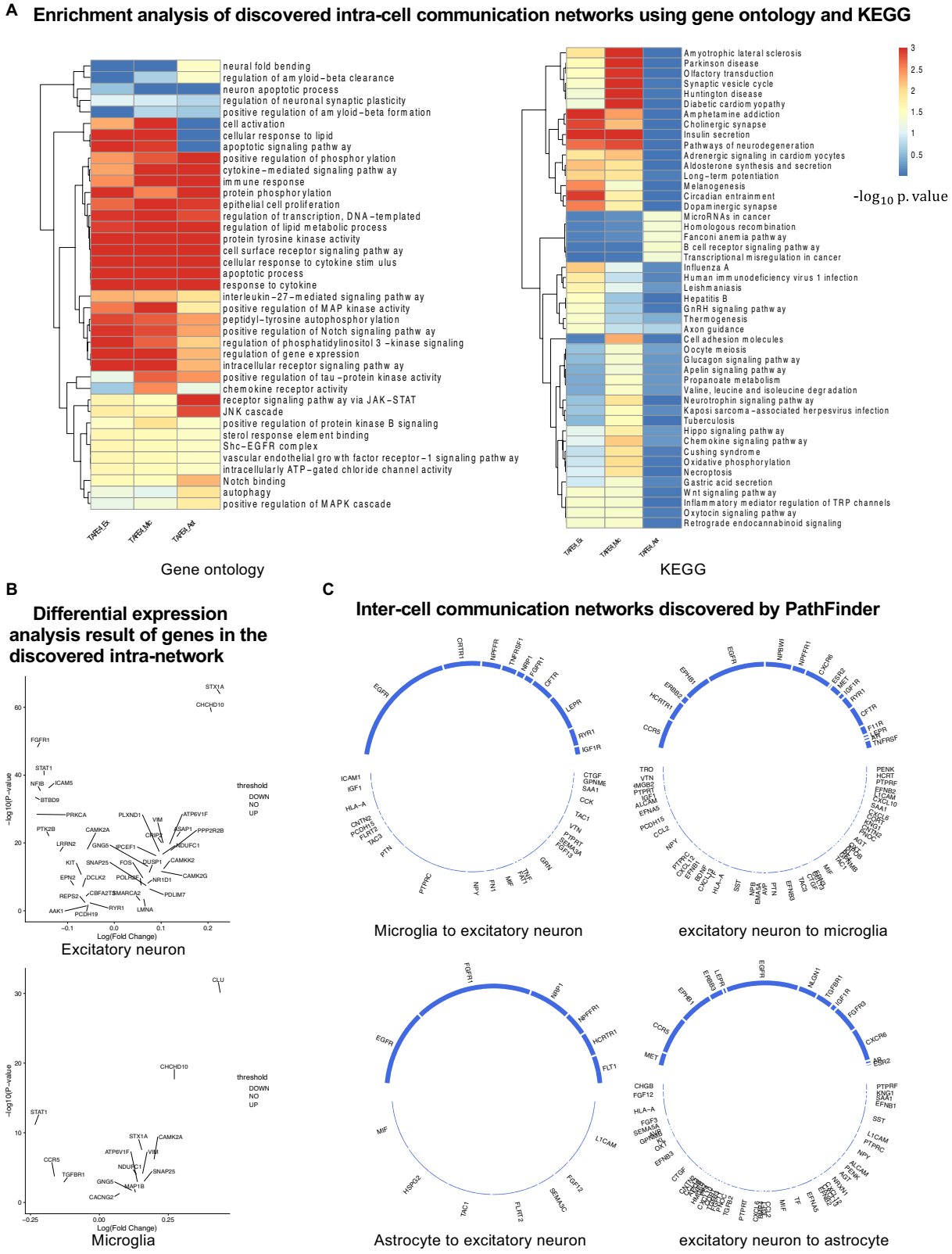

FIGURE 4
Analyses of the results. **(A)** KEGG and GO enrichment analyses on all discovered intra-networks. **(B)** Differential expression analysis. **(C)** Inter-cell communication networks. All ligands are from DEGs of the ligand cells. Receptors are marked as blue.

highly differentially expressed in the TAFE4 group, which further confirms the important role of STAT1.

Autophagy

In addition to inflammation, the *Apoptotic* and *Apoptotic signaling pathways* are enriched in the neuron and the microglia. Autophagy is a lysosome-dependent, homeostatic process, in which organelles and proteins are degraded and recycled into energy. Autophagy has been linked to Alzheimer's disease pathogenesis through its merger with the endosomal-lysosomal system, which has been shown to play a role in the formation of the latter amyloid-β plaques (Funderburk et al., 2010). One hypothesis states that irregular autophagy stimulation results in increased amyloid-β production (Yu et al., 2005). The existence of APOE4 may also affect the process of autophagy, leading to the accumulation of amyloid-β in the brain affected by AD. Particularly, CLU and FOXO1 genes are identified in the intra-network of microglia and astrocytes. CLU is one of the top AD candidate genes. Some study shows that it is a causal gene of AD-affected hippocampal connectivity (Zhang et al., 2015). Moreover, it is shown that CLU protein interacts with Aβ, reduces its aggregation, and protects against its toxic effects (Beeg et al., 2016). Many studies have shown that FOXO1 induces autophagy in cardiomyocytes and cancer cells. FOXO1 has been identified as a gene that encodes for a transcription factor involved in modulating autophagy in neurons (Xu et al., 2011).

Lipid transportation

The regulation of lipid metabolic process and cellular response to lipids are enriched in the intra-communication network of all three cell types. The enriched genes included NR1D1, EGR1, and BRCA1. It has been proved that APOE4 is involved in the lipid transportation and metabolism (Tindale et al., 2017). The existence of APOE4 in the astrocyte may disturb the brain lipid composition and thus affect the blood–brain barrier (BBB) function (Chew et al., 2020). All these results confirm the influence of APOE4 in the progress of AD and the dysfunction and death of the neuron.

JAK-STAT signaling pathway

In the intra-communication network of the astrocyte, the receptor signaling pathway via JAK–STAT is enriched with the corresponding gene: STAT3, SOCS3, HMGA2, and STAT1. The JAK–STAT signaling pathway has been reported to be the inducer of astrocyte reactivity (Ben Haim et al., 2015). The enrichment of the pathway indicates that the existence of APOE4 in astrocytes can influence the function of the JAK–STAT signaling pathway, and the pathway reversely affects the activity of the astrocyte.

Evaluation of the intra-cell signaling networks on human cirrhosis

In this section, we further evaluate the intra-cell signaling networks on human cirrhosis on endotheilal, marcrophages, and T cells. The network extraction procedure is the same as the AD cohort. The gene expression and the pathway enrichment analysis result are shown in Figure 5. The final intra-networks for each cell type are shown in Figure 6. Before the analysis, we compare the extracted intra-cell network of cirrhosis with that obtained from the AD cohort. We merge the genes from all three cell types together for AD and cirrhosis separately and then compare the common genes from both cohorts. There are 269 genes from cirrhosis and 110 genes from AD. However, there are only 14 common genes, which demonstrate that PathFinder is disease- and expression-specific. We further explore the networks identified by the PathFinder model and their relationship with cirrhosis.

The role of immune cells in liver diseases

Immune cells and various signaling pathways play an important role in the pathogenesis of liver diseases. Gene CCR9 is activated in the intra-cell signaling network of both endothelial and T cells. Studies have found that, in a mouse model of NASH, the CCR9/CCL25 axis promotes the recruitment of macrophages and the formation of fibrosis, providing a new potential therapeutic target for NASH (Morikawa et al., 2021). On the other hand, liver NKT cells accumulate in a CXCR6-dependent manner early after injury, exacerbating the inflammatory response and promoting the progression of liver fibrosis, suggesting that the CXCR6/CXCL16 pathway may be an effective target for the treatment of liver fibrosis (Wehr et al., 2013). CXCR6 is discovered by PathFinder for the intra-cell signaling network of both endothelial and macrophages, which further confirms it. Additionally, β-arrestin1 (ARRB1) activated at the signaling network of all three cell types was reported to interact with pro-GDF15, promoting its cleavage and maturation in the Golgi apparatus, and the absence of ARRB1 significantly exacerbates hepatic steatosis, fibrosis, and inflammation (Zhang et al., 2020).

Liver fibrosis and its reversibility

The development of liver fibrosis is a complex and potentially reversible process. In its early stages, liver fibrosis may not immediately present severe symptoms but can eventually progress to cirrhosis and affect multiple organs. CREB is a highly activated gene discovered by PathFinder. Research has found that CREB, a molecule downstream of the cAMP signaling pathway, can serve as a therapeutic target for fibrosis (Li et al., 2019). Furthermore, insulin-like growth factor 1 (IGF1) and its receptor IGF1R play a crucial role in liver health and function, primarily expressed in the liver tissue. Studies on liver fibrosis have revealed the core role of the IGF1/IGF1R signaling system in controlling the liver fibrosis process (Gui et al., 2023). In the intra-cell signaling network of all three cell types identified by PathFinder, IGF1R is activated and further triggers target GNLY and HBEGF through FGFR3. Although there is not enough literature discussing their relationship with cirrhosis, exploiting the molecular mechanisms and functionality may provide new insights into studying cirrhosis and be helpful in developing more effective treatments to solve liver disease problems.

Liver disease transition process

In the intra-cell signaling networks identified by PathFinder, genes EGR1 and ERBB3 are highly activated. In the liver disease transition process, chronic hepatitis and cirrhosis are major factors leading to the majority of hepatocellular carcinomas (HCC). Concurrently, non-alcoholic fatty liver disease (NAFLD) has become a global epidemic, not only associated with the development of metabolic syndrome but also regarded as a pathway leading to severe liver diseases such as cirrhosis and hepatocellular carcinoma. In this transition process, EGR1 has been discovered as a key regulator of NAFLD, presenting potential as a potent target for intervening in

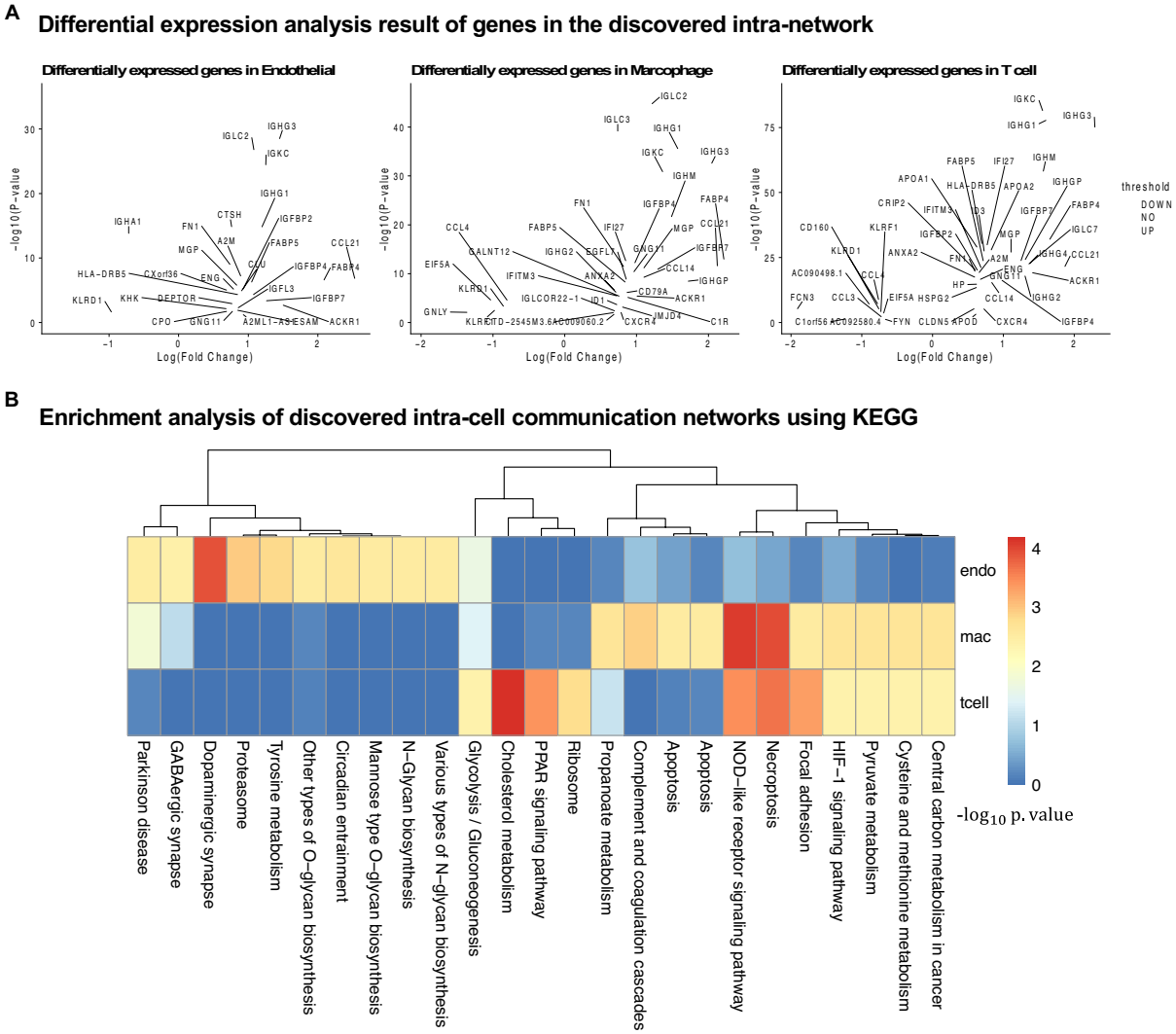

FIGURE 5
Analysis of the results for the cirrhosis cohort. **(A)** Differential expression analysis for all three cell types. **(B)** KEGG pathway enrichment analysis using the intra-cell networks discovered by PathFinder.

NAFLD (Guo et al., 2023). Additionally, research has identified ERBB3 as a potential serum marker for early HCC in patients with chronic hepatitis and cirrhosis (Nasiri et al., 2020). A deeper understanding of the mechanisms underlying liver disease transition will provide insights into therapeutic strategies for related diseases.

Core multi-cell inter-cell communication networks associated with the APOE4 genotype

To further understand the complex signaling flow and mechanism behind the APOE4 and AD pathology, we further generate inter-cell communication networks between three different cell types using PathFinder, as shown in Figure 4C. First, we can see that, compared to astrocytes, microglia have much more interactions with neurons. This may indicate that the existence of APOE4 in the astrocyte may activate the functionality of microglia and then cause abnormal activities in the neurons. Among all interactions, several interesting interactions appealed to the result. First, the MIF secreted by the astrocyte interacts with the EGFR in the neuron and follows downstream signaling. The MIF is a well-known proinflammatory cytokine that promotes the production of other immune mediators. Increased expression of MIF can contribute to chronic neuroinflammation and neurodegeneration (Tavassoly et al., 2020). EGFR is a potential target for treating AD-induced memory loss (Zhu et al., 2011; Wang et al., 2012). The increased expression level of MIF could be the signature of activated astrocytes, and the MIF further triggers the expression of EGFR and the subsequent downstream network in the neuron, which contributes to neuron inflammation and degeneration.

In addition to MIF in astrocytes, many ligands for receptor EGFR are also identified in microglia, including ICAM1, IGF1, HLA-A, CNTN2, PCDH15, FLRT2, TAC3, PTN, and PTPRC. The downregulation of PTPRC is reported to contribute to the overproduction of Aβ and neuron loss (Brito-Moreira et al., 2017). Another interaction is the *NLGN1* gene which is expressed in neurons that interact with the *NRXN1* gene in the astrocyte. The amyloid-β oligomers are synaptotoxins that build up in the brains of patients

10.3389/fncel.2024.1369242

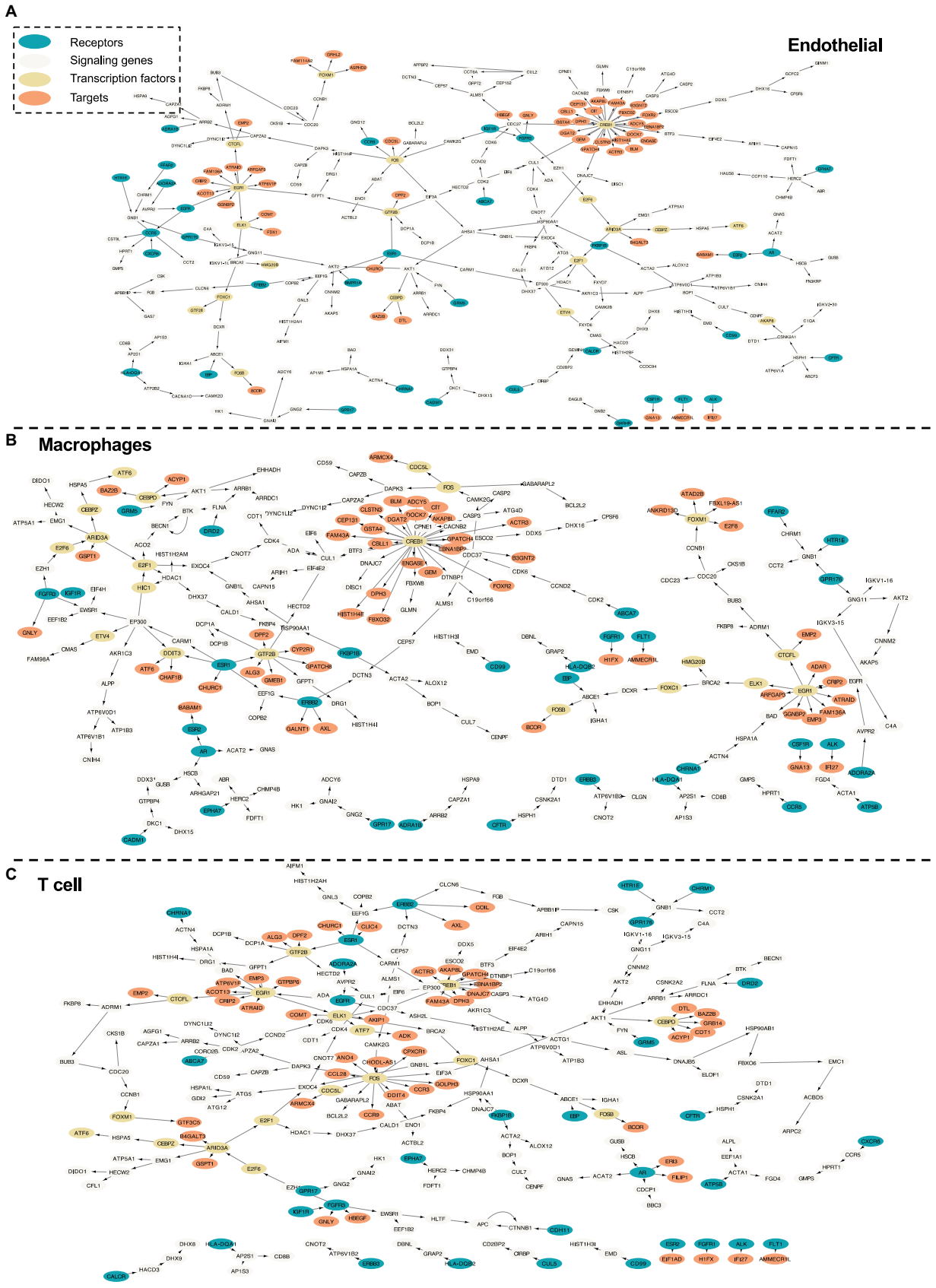

FIGURE 6
Intra-cell communication networks discovered by the PathFinder model for the human cirrhosis cohort. **(A)** Endothelial; **(B)** Macrophages; **(C)** T cell.

and are thought to contribute to the memory impairment in AD. It has been shown that the interaction of neurexins (Nrxs) and neuroligins (NLs) is critical for synapse structure, stability, and function (Tyzack et al., 2017). The dysregulation of the interaction between Nrxs and NLs may contribute to the formation of amyloid-β oligomer. The *EFNA5* in the neuron is upregulated in the neuron and interacts with *EPHB1* and downstream *STAT3* signaling in the astrocyte. This interaction is closely related to the ephrin-B1-mediated stimulation. The analysis has shown that the ephrin-B1-mediated stimulation induces a protective and anti-inflammatory signature in astrocytes and can be regarded as "help-me" signal of neurons that failed in early amyotrophic lateral sclerosis (ALS) (Lambert et al., 2018). Such signals could also play an important role in triggering inflammation and neuron degeneration in the CNS system.

Conclusion and discussion

In this study, we propose PathFinder, which is the first deep-learning model with a graph transformer that can be used to extract both intra- and inter-cell communication networks using scRNA-seq data. Through a case study using an AD scRNA-seq dataset from mice, we evaluate the effectiveness of PathFinder from multiple perspectives. First, the quantitative analysis confirms that PathFinder performs well in separating cells from different conditions by leveraging the difference of expression patterns in the signaling paths. Furthermore, the learned path score is robust and consistent in repeat runs. We further evaluate the correctness of extracted networks through extensive literature searches. The resulting network aligned well with many recent discoveries on the AD pathology, which further proved the effectiveness of the proposed PathFinder. Additionally, the current version of PathFinder has a few potential limitations to be improved in the future studies. First, it requires many samples in training to produce reasonable results. Second, it relies on the pre-defined paths from the database to learn and extract meaningful patterns and is unable to discover new signaling flows. Third, currently, it is hard to validate the discovered signaling pathway quantitatively as there is no existing benchmark for conducting this process. All these limitations warrant further investigation. For example, we can construct a common benchmark to evaluate the performance of all signaling network inference methods quantitatively. We will also improve the model in our future work.

Methodology

Gene-gene interaction database collection and processing

To construct the gene–gene interaction database, the raw interaction data were collected from NicheNet software (Browaeys et al., 2020). The raw interaction data were divided into three types: ligand-receptor network, signaling network, and gene-regulation network. The original network contained 12,019 interactions/1,430 genes, 12,780 interactions/8,278 genes, and 11,231 interactions/8,450 genes, respectively. To construct the intra- and inter-network database, the data were further processed by the following steps.

First, ligands and receptors were collected by gathering the source and target of the ligand-receptor network. There were a total of 688 ligands and 857 receptors. Then, interactions in the ligand-receptor network were divided into two types. If one interaction exists in both directions in the database, we labeled it as bidirectional. Otherwise, we labeled it as directional. After processing, there were 11,880 directional interactions and 139 bidirectional interactions.

The gene-regulation network was processed as follows. First, 1,639 transcriptional factors (TFs) were collected (Fan et al., 2021). For convenience, TFs that exist in either the ligand or receptor list were removed. Finally, 1,632 TFs were collected. Then, three different types of regulation were collected in the gene-regulation interaction network, which are ligand regulation, receptor regulation, and TF regulation. To label each interaction into one of three types, all the interactions in the network were removed if the source gene was not in the ligand, receptor, or TF list. Then, the interactions were labeled based on the type of source (e.g., if the source of interaction is a receptor, we label it as receptor regulation). After processing, there were 1,329 ligand-regulation interactions, 272 receptor-regulation interactions, and 6,706 TF-regulation interactions.

Finally, the signaling network was processed as follows. First, all interactions were removed if they existed in either the ligand-receptor or the gene-regulation network. Then, the interactions were further divided into receptor-TF, receptor-signaling, signaling-TF, and signaling-signaling. To be more specific, if the source of interaction is in the receptor list and the target of interaction is in the TF list, the interaction was labeled as receptor-TF. If the source of interaction is in the receptor list and the target is not in the tTF list, the interaction was labeled as receptor-signaling. If the source of interaction is not in the receptor list and the target of interaction is in the TF list, the interaction was labeled as signaling-TF. If neither the source nor target of interaction is in the TF and receptor lists, the interaction was labeled as signaling-signaling. The interactions that cannot be classified into one of the specified groups were removed for convenience. Finally, there are 31 receptor-TF interactions, 524 receptor-signaling interactions, 975 signaling-TF interactions, and 9,745 signaling-signaling interactions.

Notations and terminologies

Terminologies

An embedding or a representation is a vector of size R^d that represents an entity, such as a gene or a path. The input embedding is the embedding input to the model, the hidden embedding is the embedding output by the middle layers of the model, and the output embedding is the embedding output by the model. With the final output embedding for an entity, we can do the classification or regression by passing it to a logistic regression or linear regression layer. An encoding is a function that transforms an entity to the embedding. Typically, the goal of a deep learning or machine learning model is to learn a model that can take the input embedding of the entity we want to predict and output the output embedding which is more reliable and powerful for the prediction. A single neural network layer will contain one or multiple trainable weight matrices. These matrices is responsible for transforming the input embedding into the output embedding. They will be updated and refined by the backward propagation and gradient descent used in the neural network.

Notations

A gene graph is denoted as $G=(V,E)$, where V is the set of gene nodes with $|V|=n$, E is the set of edges and $E\subseteq V\times V$. The node embedding set is denoted by $X=[x_1,x_2,\ldots,x_n]^T\in R^{n\times d}$, where $x_u\in R^d$ is the embedding vector of the node u. The graph structure is defined by an adjacency matrix $A\in[0,1]^{n\times n}$, where $A_{uv}=1$ indicate there is an edge from the node u to node v and $A_{uv}=0$ otherwise. Furthermore, a set of paths sampled from a graph is denoted as $P=\{p_1,p_2,\ldots,p_p\}$, where p_m is the m-th path, which is a list to store the nodes of the path in order. Paths can have different lengths, and we denote the length of path m be l_m.

Preliminary of transformer and Graphormer

The transformer is a powerful architecture in the deep learning field. It consists of multiple transformer layers. Each transformer layer has two parts: a multi-head self-attention and a point-wise feed-forward network (FFN) with residual connection applied between each part. Let $H^{l-1}\in R^{n\times h_{emb}}$ be the embedding of nodes in layer $l-1$, and H_u^{l-1} is the embedding of the node u in layer $l-1$, the computation of multi-head self-attention is:

$$Q^{l,i}=H^{l-1}W_Q^{l,i},K^l=H^{l-1}W_K^{l,i},V^l=H^{l-1}W_V^{l,i},$$

$$head_i=Attention\left(Q^{l,i},K^{l,i},V^{l,i}\right)=SoftMax\left(\frac{Q^{l,i}K^{l,iT}}{\sqrt{d_k}}\right)V^{l,i},$$

$$O^l=Concat\left(head_1,\ldots,head_h\right)W_O^l,$$

where $W_Q^{l,i},W_K^{l,i},W_V^{l,i}\in R^{h_{emb}\times d_k}$, and $W_O^l\in R^{hd_k\times h_{emb}}$ are all trainable weight matrix, h is the number of heads, $O^l\in R^{n\times h_{emb}}$ is the output from the multi-head self-attention in layer l, *Concat* is the concatenation function to combine multiple vectors into one single large vector. For simplicity, we let $h\times d_k=h_{emb}$. The output O^l will then be fed into a point-wise feed-forward network. The computation of the point-wise feed-forward network is:

$$FFN(x)=ReLu\left(xW_1^l+b_1^l\right)W_2^l+b_2^l,$$

where $W_1^l\in R^{h_{emb}\times h_{emb}},W_2^l\in R^{h_{emb}\times h_{emb}}$, $b_1^l\in R^{2h_{emb}}$, and $b_2^l\in R^{h_{emb}}$ are all trainable weight matrix and bias. Notice that here we slightly modify the hidden size of the feed-forward network of the original model. The embedding of each node $O_i^l\in R^{h_{emb}}$ will be input into this FFN for further processing.

However, the vanilla transformer cannot be used directly on the graph structure data as it lacks a critical part for encoding the topological information into the model. To deal with this issue, Graphormer proposed several novel encodings into the model. Specifically, they introduced centrality encoding, spatial encoding, and edge encoding. The centrality encoding is used to embed the graph centrality information into the model. Given the input data X, the computation of centrality encoding is:

$$H^0=X+Z^-\left\{deg^-(G)\right\}+Z^+\left\{deg^+(G)\right\},$$

where the Z^-,Z^+ are all trainable embedding vectors and $deg^-(G),deg^+(G):G\to R^n$ are the function to compute the in-degree and out-degree of each node in the graph G. The spatial and edge encoding is used to encode the graph structure into the model. With the spatial and edge encoding, the self-attention is revised as:

$$head_i=SoftMax\left(\frac{Q^{l,i}K^{l,iT}}{\sqrt{d_k}}+b_i\left\{\phi(G)\right\}+c_i\right)V^{l,i},$$

where b^i is trainable embedding vectors to encode the spatial information at head i and $\phi(G):G\to R^{n\times n}$ is the function to compute the shortest path length between each two nodes. If two nodes are not connected, a special value will be used. $c^i\in R^{n\times n}$ is the edge embedding and $c_{uv}^i=\frac{1}{N}\sum_{n=1}^{N}x_{en}w_n^{iT}$, where x_{en} is the edge feature of the n-th edge in the shortest path between node u and node v and the w_n^i is trainable weight vector of n-th edge of head i. Note that both the spatial and edge encodings are unique across different layers.

Architecture of PathFinder

The PathFinder model consists of three components, namely, the node encoder, path encoder, and graph encoder. The overall architecture of the PathFinder model is shown in Figure 1, lower. The rationale behind PathFinder is that, if a model can identify disease cells from normal cells, it must learn useful knowledge from the gene expression profile to help it make that prediction. In PathFinder, we introduce the path encoder to let the model make the prediction based on the importance of the signaling paths with their corresponding expression. In this way, if the model can make a reasonable prediction, it must have the ability to distinguish differential expressed signaling paths from the other paths, and that is exactly what we are looking for. Furthermore, since the paths are pre-defined from the physical interaction database in a biologically meaningful way, the extracted signaling paths are inherently biologically meaningful. PathFinder can be seen as a simulator to simulate the signaling path in the cell and use it to make the prediction. Below, we discuss each component in detail.

Node encoder

The architecture of the node encoder is similar to the Graphormer, which stacks L transformer layer with centrality encoding, spatial encoding, and edge encoding. The input to PathFinder is the expression value of each gene in a cell sample. However, we made several modifications to the original architecture. First, the hidden size in the point-wise feed-forward network is all h_{emb} in both two layers for simplicity. Second, the edge encoding in PathFinder is modified. In the original

Graphormer, the edge encoding is computed by all the edges in the shortest path between two nodes, which can capture long-range information in the graph. However, the localized feature in the graph will be smoothed in such a manner. Instead, PathFinder aims for the node embedding learned from the node encoder to focus on the localized information in the graph. Therefore, direct edge encoding is proposed. The direct edge encoding is computed by:

$$c_{uv}^{i} = x_{uv} w^{iT}$$

Where x_{uv} is the edge feature of the edge between node u and node v. If there is not an edge between two nodes, the direct edge encoding is set to a special vector for simplicity. By doing this, the node encoder becomes adept at learning node embedding that capture localized information. Finally, the spatial encoding is also revised in PathFinder. Since here the graph structure is identical for all samples and the node order invariant is automatically held, we can learn a specific spatial encoding for each pair of two nodes. Therefore, we design the node index encoding in the PathFinder model. The node index encoding is not computed from the length of the shortest path between each pair of nodes but is directly learned for each pair of two genes, namely, for each pair of two genes, a unique encoding is learned for each head in each layer of the node encoder.

Path encoder

Furthermore, the path encoder is responsible for learning gene signaling path embedding, utilizing the node embedding in the graph and the pre-defined path list of the graph. The details of the pre-defined path list are illustrated below. Suppose there are p unique paths in the path list P, where the length of the m-th path is l_m and the total number of nodes in the path list is k (count repeated nodes in different paths). Denote the node embedding output from the layer l as H^l, we first learn a path-specific embedding through:

$$U_i^l = scatter\left(H^l\right)_i W_u^l + b_u^l,$$

where $W_u^l \in R^{h_{emb} \times u}$ and $b_u^l \in R^u$ are all trainable weight matrix, *scatter* is a function to reorder and scatter the node in the graph into the order of the pre-defined path list. For example, suppose there are five embedding genes output from the node encoder. That is $H^l \in R^{5 \times h_{emb}}$. We label each gene from 1 to 5. Suppose there are two paths. The first path is 1->3->4. The second path is 2->3->4->5. Then, the $scatter\left(H^l\right)$ will output a new matrix with the size of 7 and each row represents a gene in a path. For instance, the third row is H_4^l since it is the third gene in the first path. $U^l \in R^{k \times u}$ is the learned path-specific embedding. For convenience, we denote $U_{m,i}^l$ as the embedding of i-th node in the m-th path. Then, path positional and path edge encodings are introduced to encode additional information for all paths. Let $\bar{U}^l$ be the result embedding after the special encodings. We have:

$$\bar{U}_{m,i}^l = U_{m,i}^l + p_i^l + e_{i,i+1}^l,$$

Where p_i^l is the learnable positional encoding vector and its value only depends on the position i, $e_{i,i+1}^l$ is the learnable edge encoding to encode the edge type between i-th node and $i+1$-th node. Then, the score of each node within the path is computed by:

$$S_{m,i}^l = \tanh\left(\bar{U}_{m,i}^l W_{s1}^l + b_{s1}^l\right) W_{s2}^l + b_{s2}^l$$

$$\bar{S}^l = ScatterSoftMax\left(S^l\right),$$

where $W_{s1}^l \in R^{u \times r}, b_{s1}^l \in R^r, W_{s2}^l \in R^{r \times r}$, and $b_{s2}^l \in R^r$ are all trainable parameters. *ScatterSoftmax* is the softmax function working within each path. The $S^l \in R^{k \times r}$ is the final r set important score for each node in each path. We let $r \times u = h_{emb}$ for simplicity. After we obtain S^l, the path embedding is computed by:

$$P^l = Flatten\left(ScatterSum\left(S^l * \bar{U}^l\right)\right)$$

$*$ is the point-wise product working on each set of important scores. That is, for each set of important scores, we do a point-wise product of that set of scores and U^l, which results in total r sets. The *ScatterSum* function is the summation on each path. *Flatten* is the function to flatten the embedding of all sets. $P^l \in R^{p \times h_{emb}}$ is the final path embedding in the layer l.

Graph encoder

In the original Graphormer, the graph embedding is learned by introducing a special node and letting it connect to all the nodes in the graph. After forwarding, the embedding of that special node is regarded as the graph embedding for the graph-level task. In PathFinder, our goal is to learn the graph embedding from the path embedding. Meanwhile, we aim to extract the important paths from the model after training it for the graph-level task. To simultaneously achieve both goals, the graph encoder is proposed. The graph encoder consists of two parts. The first part is a trainable path weight and the sigmoid function to assign each path with different scores. The second part is the jumping knowledge network to combine the graph embedding in each layer and compute the final embedding.

In PathFinder, the graph embedding is learned by integrating all the path embeddings from each layer, which requires an important score for each path. Normally, the score is computed based on one sample. However, such a score is not robust and may vary a lot even with a minor variation of the path embedding (Xu et al., 2018; Chen et al., 2019; Fan et al., 2021). To avoid the issue and learn a robust important score across the whole dataset, the trainable path score $M \in R^p$ is introduced. M is identical to all samples and layers and learned through backpropagation. The path important score is computed by:

$$I = Sigmoid\left(M\right),$$

where $I \in R^p$ is the important score for each path. Then, the graph embedding of layer l is computed by:

$$g^l = IP^l,$$

where g^1 is the graph embedding of layer l. The final step of the graph encoder is to integrate the graph embedding of each layer and learn a final embedding. Here, we utilize the idea of JumpingKnowledge network (Xu et al., 2018) and compute the final graph embedding by:

$$G = MaxPooling\left(Concat\left(g^1, g^2, \ldots, g^L\right)\right),$$

where *MaxPooling* is the max pooling function and $G \in R^{h_{emb}}$ is the final graph embedding learned by PathFinder. Finally, the graph embedding is used to classify the cell sample into the corresponding condition (control/test). The prediction is a typical binary prediction computed by:

$$p = SoftMax\left(GW_p\right),$$

Where $W_p \in R^{h_{emb} \times 2}$ is the trainable projection matrix and p is the predicted distribution.

Training and regularization of PathFinder

To train the PathFinder model, the negative log-likelihood (NLL) loss is applied. Let the p_i^c be the predicted probability of the true condition of cell i, then the NLL loss is computed by:

$$\mathcal{L}_{class} = \sum_{i=1}^{N} -\log\left(p_i^c\right)$$

Where the N is the number of cells in the dataset. Meanwhile, to regularize the training of the model and learn biological meaningful paths from the model, the regularization term is introduced to the path score M. Intuitively, the path that has a higher total fold change should have a higher path score. Furthermore, we designed three different regularization terms to generate different important paths by introducing the prior path score. Specifically, these three regularizations are upregulated path, downregulated path, and differentially expressed path regularization. Let the fc_i^m be the log fold change of gene j in path m, then the prior path score is computed by:

$$S_{up}^m = Normalization\left(mean\left(\sum_j fc_j^m\right)\right)$$

$$S_{down}^m = Normalization\left(mean\left(\sum_j -fc_i^m\right)\right)$$

$$S_{\deg}^m = Normalization\left(mean\left(\sum_j |fc_i^m|\right)\right)$$

Where the S_{up}^m, S_{down}^m, and $S_{\deg}^m$ are the prior path scores for upregulated, downregulated, and differential expressed regularization, respectively. *Normalization* is the min-max normalization across all paths. Suppose we use the upregulated prior score, the regularization loss is computed by:

$$\mathcal{L}_{reg} = D_{KL}\left(I \parallel S_{up}^m\right)$$

The final loss is:

$$\mathcal{L} = \mathcal{L}_{class} + \beta \mathcal{L}_{reg}$$

Where β is the weight of the regularization term.

Predefined path list

To train the PathFinder model, the path list needs to be defined before the training. Given the collected gene–gene interaction database and the input variable gene list, we designed several choices to generate a predefined path list. The first choice is the shortest path. For this choice, the shortest path between each pair of genes in the dataset will be computed and collected given the gene–gene interaction network. The second choice is to generate all the possible paths that start from the receptor and end in the target, which can also be performed using the gene–gene interaction database. To constrain the path, the minimum length of the path is set to be 3 unless the path is a receptor regulation interaction. The maximum length of the path is set to be 10.

Experimental details

We conduct experiments to validate the effectiveness of PathFinder on TAFE_ex, TAFE_mic, and TAFE_ast cell sample datasets. For each dataset, we randomly split datasets into train/validation/test sets with a ratio of 0.7/0.1/0.2. We train the model using the train set and validate the performance of the model using the validation set. Finally, we save the model that achieves the best performance on the validation set and report the performance of the saved model on the test set. We use the area under the curve (AUC) as the performance metric for selecting the best model. We repeat experiments on each dataset five times (with a different random split applied to the dataset each time) and report the mean results and the standard deviation. The model and training hyperparameters are described as follows: We set the number of layers as 6 and the hidden size h_{emb} as 128. The number of heads and scores set r as 8. For each experiment, we set the number of training epochs as 30, the learning rate as

0.0005, the dropout rate as 0.1, the regularization weight β as 0.1 for TAFE_ex and TAFE_mic, and 1.0 for TAFE_ast.

Generation of the intra- and inter-cell communication network

After the PathFinder model is trained, the generation of an intra-cell communication network is straightforward. Concretely, we first average the path weight learned from five repeated experiments to get the final path weights. Furthermore, the top K paths are extracted and combined to generate the intra-cell communication network. The generation of the inter-cell communication network is as follows. Let the cell that provides ligands be the ligand cell and the cell providing receptors be the receptor cell. The intra-cell communication network is first generated. Then, the ligands of the ligand cell and receptors of the receptor cell will be extracted from their respective intra-networks. Then, the ligand–receptor pairs are selected given the ligand–receptor database. Finally, the kept pairs will be linked and the inter-network is generated.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164507. The source code of PathFinder is publicly accessible on github: https://github.com/fuhaililab/PathFinder.

Author contributions

JF: Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. MP: Writing – review & editing. GL: Writing – review & editing. PP: Writing – review & editing. YC: Methodology, Writing – review & editing. FL: Conceptualization, Funding acquisition, Methodology, Writing – original draft, Writing – review & editing. HS: Writing – original draft, Formal analysis, Validation.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was partially supported by NIA R56AG065352 (to Li), 1R21AG078799-01A1 (to Li), and 1RM1NS132962-01 (to Dickson/Sardiello/Cooper/Li).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncel.2024.1369242/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Additional evaluation results. **(A)** Comparison of the differential expression level between paths identified by PathFinder and the rest in cirrhosis cohort. **(B)** The learned path scores of PathFinder on different runs on cirrhosis cohort.

References

Akiyama, H., Barger, S., Barnum, S., Bradt, B., Bauer, J., Cole, G. M., et al. (2000). Inflammation and Alzheimer's disease. *Neurobiol. Aging* 21, 383–421. doi: 10.1016/S0197-4580(00)00124-X

Beeg, M., Stravalaci, M., Romeo, M., Carrá, A. D., Cagnotto, A., Rossi, A., et al. (2016). Clusterin binds to Aβ1–42 oligomers with high affinity and interferes with peptide aggregation by inhibiting primary and secondary nucleation*. *J. Biol. Chem.* 291, 6958–6966. doi: 10.1074/jbc.M115.689539

Ben Haim, L., Ceyzériat, K., Sauvage, M. A. C.-d., Aubry, F., Auregan, G., Guillermier, M., et al. (2015). The JAK/STAT3 pathway is a common inducer of astrocyte reactivity in Alzheimer's and Huntington's diseases. *J. Neurosci.* 35, 2817–2829. doi: 10.1523/JNEUROSCI.3516-14.2015

Brito-Moreira, J., Lourenco, M. V., Oliveira, M. M., Ribeiro, F. C., Ledo, J. H., Diniz, L. P., et al. (2017). Interaction of Aβ oligomers with Neurexin 2α and Neuroligin 1 mediates synapse damage and memory loss in mice. *J. Biol. Chem.* 292, 7327–7337. doi: 10.1074/jbc.M116.761189

Browaeys, R., Saelens, W., and Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* 17, 159–162. doi: 10.1038/s41592-019-0667-5

Butturini, E., Cozzolino, F., Boriero, D., Carcereri de Prati, A., Monti, M., Rossin, M., et al. (2018). S-glutathionylation exerts opposing roles in the regulation of STAT1 and STAT3 signaling in reactive microglia. *Free Radic. Biol. Med.* 117, 191–201. doi: 10.1016/j.freeradbiomed.2018.02.005

Cai, D., and Lam, W. *Graph transformer for graph-to-sequence learning. In AAAI* (2020).

Chen, J., Wu, X., Rastogi, V., Liang, Y., and Jha, S. Robust attribution regularization. In: NeurIPS (2019).

Chew, H., Solomon, V. A., and Fonteh, A. N. (2020). Involvement of lipids in Alzheimer's disease pathology and potential therapies. *Front. Physiol.* 11:598. doi: 10.3389/fphys.2020.00598

Choi, H., Sheng, J., Gao, D., Li, F., Durrans, A., Ryu, S., et al. (2015). Transcriptome analysis of individual stromal cell populations identifies stroma-tumor crosstalk in mouse lung Cancer model. *Cell Rep.* 10, 1187–1201. doi: 10.1016/j.celrep.2015.01.040

Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. (2020). CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* 15, 1484–1506. doi: 10.1038/s41596-020-0292-x

Fan, W., Jin, W., Liu, X., Xu, H., Tang, X., Wang, S., et al. (2021). Jointly attacking graph neural network and its explanations. *ArXiv abs/2108.03388*. doi: 10.48550/arXiv.2108.03388

Feng, J., Zeng, A., Chen, Y., Payne, P., and Li, F. (2020). Signaling interaction link prediction using deep graph neural networks integrating protein-protein interactions and omics data. *bioRxiv 2020.12.23.424230*. doi: 10.1101/2020.12.23.424230

Funderburk, S. F., Marcellino, B. K., and Yue, Z. (2010). Cell 'self-eating' (autophagy) mechanism in Alzheimer's disease. *Mt. Sinai J. Med.* 77, 59–68. doi: 10.1002/msj.20161

Gui, R., Li, W., Li, Z., Wang, H., Wu, Y., Jiao, W., et al. (2023). Effects and potential mechanisms of IGF1/IGF1R in the liver fibrosis: a review. *Int. J. Biol. Macromol.* 251:126263. doi: 10.1016/j.ijbiomac.2023.126263

Guo, Y., Miao, X., Sun, X., Li, L., Zhou, A., Zhu, X., et al. (2023). Zinc finger transcription factor Egf1 promotes non-alcoholic fatty liver disease. *JHEP Rep.* 5:100724. doi: 10.1016/j.jhepr.2023.100724

Halliday, G., Robinson, S. R., Shepherd, C., and Kril, J. (2000). Alzheimer's disease and inflammation: a review of cellular and therapeutic mechanisms. *Clin. Exp. Pharmacol. Physiol.* 27, 1–8. doi: 10.1046/j.1440-1681.2000.03200.x

Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Inductive representation learning on large graphs. *arXiv:1706.02216v4*. doi: 10.48550/arXiv.1706.02216

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M. III, Zheng, S., Butler, A., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell* 184, 3573–3587.e29. doi: 10.1016/j.cell.2021.04.048

Hu, Z., Dong, Y., Wang, K., and Sun, Y. (2020). "Heterogeneous graph transformer" in *Proceedings of the web conference 2020 2704–2710 (Association for Computing Machinery)* (New York, NY).

Hu, X., Herrero, C., Li, W. P., Antoniv, T. T., Falck-Pedersen, E., Koch, A. E., et al. (2002). Sensitization of IFN-γ Jak-STAT signaling during macrophage activation. *Nat. Immunol.* 3, 859–866. doi: 10.1038/ni828

Hu, Y., Peng, T., Gao, L., and Tan, K. (2021). CytoTalk: De novo construction of signal transduction networks using single-cell transcriptomic data. *Sci. Adv.* 7:eabf1356. doi: 10.1126/sciadv.abf1356

Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* 50, 1–14. doi: 10.1038/s12276-018-0071-8

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58, 610–620. doi: 10.1016/j.molcel.2015.04.005

Lambert, S. A., Jolma, A., Campitelli, L. F., das, P. K., Yin, Y., Albu, M., et al. (2018). The human transcription factors. *Cell* 172, 650–665. doi: 10.1016/j.cell.2018.01.029

Li, G., Jiang, Q., and Xu, K. (2019). CREB family: a significant role in liver fibrosis. *Biochimie* 163, 94–100. doi: 10.1016/j.biochi.2019.05.014

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., et al. (2019). Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* 570, 332–337. doi: 10.1038/s41586-019-1195-2

Morikawa, R., Nakamoto, N., Amiya, T., Chu, P. S., Koda, Y., Teratani, T., et al. (2021). Role of CC chemokine receptor 9 in the progression of murine and human non-alcoholic steatohepatitis. *J. Hepatol.* 74, 511–521. doi: 10.1016/j.jhep.2020.09.033

Nasiri, E., Sankowski, R., Dietrich, H., Oikonomidi, A., Huerta, P. T., Popp, J., et al. (2020). Key role of MIF-related neuroinflammation in neurodegeneration and cognitive impairment in Alzheimer's disease. *Mol. Med.* 26:34. doi: 10.1186/s10020-020-00163-5

Rogers, J., Webster, S., Lue, L. F., Brachova, L., Harold Civin, W., Emmerling, M., et al. (1996). Inflammation and Alzheimer's disease pathogenesis. *Neurobiol. Aging* 17, 681–686. doi: 10.1016/0197-4580(96)00115-7

Rong, Y., et al. (2020). Self-supervised graph transformer on large-scale molecular data. *arXiv:2007.02835v2*. doi: 10.48550/arXiv.2007.0283

Saint-Antoine, M. M., and Singh, A. (2019). Network inference in systems biology: Recent developments, challenges, and applications. *Curr. Opin. Biotechnol.* 63, 89–98. doi: 10.1016/j.copbio.2019.12.002

Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. *Nature* 541, 331–338. doi: 10.1038/nature21350

Tavassoly, O., Sato, T., and Tavassoly, I. (2020). Inhibition of brain EGFR activation: a novel target in neurodegenerative diseases and brain injuries. *Mol. Pharmacol.* 98, 13–22. doi: 10.1124/mol.120.119909

Tindale, L. C., Leach, S. R., Spinelli, J. J., and Brooks-Wilson, A. R. (2017). Lipid and Alzheimers disease genes associated with healthy aging and longevity in healthy oldest-old. *Oncotarget* 8, 20612–20621. doi: 10.18632/oncotarget.15296

Tyzack, G. E., Hall, C. E., Sibley, C. R., Cymes, T., Forostyak, S., Carlino, G., et al. (2017). A neuroprotective astrocyte state is induced by neuronal signal EphB1 but fails in ALS models. *Nat. Commun.* 8:1164. doi: 10.1038/s41467-017-01283-z

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *arXiv:1706.03762v7*. doi: 10.48550/arXiv.1706.03762

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y., et al. (2017). Graph attention networks. *arXiv:1710.10903v3*. doi: 10.48550/arXiv.1710.10903

Wang, L., Chiang, H. C., Wu, W., Liang, B., Xie, Z., Yao, X., et al. (2012). Epidermal growth factor receptor is a preferred target for treating amyloid-β–induced memory loss. *Proc. Natl. Acad. Sci.* 109, 16743–16748. doi: 10.1073/pnas.1208011109

Wang, C., Xiong, M., Gratuze, M., Bao, X., Shi, Y., Andhey, P. S., et al. (2021). Selective removal of astrocytic APOE4 strongly protects against tau-mediated neurodegeneration and decreases synaptic phagocytosis by microglia. *Neuron* 109, 1657–1674.e7. doi: 10.1016/j.neuron.2021.03.024

Wehr, A., Baeck, C., Heymann, F., Niemietz, P. M., Hammerich, L., Martin, C., et al. (2013). Chemokine receptor CXCR6-dependent hepatic NK T cell accumulation promotes inflammation and liver fibrosis. *J. Immunol.* 190, 5226–5236. doi: 10.4049/jimmunol.1202909

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K., and Jegelka, S. (2018). Representation Learning on Graphs with Jumping Knowledge Networks. *ArXiv*.

Xu, P., Das, M., Reilly, J., and Davis, R. (2011). JNK regulates FoxO-dependent autophagy in neurons. *Genes Dev.* 25, 310–322. doi: 10.1101/gad.1984311

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. *How powerful are graph neural networks?* (2018).

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K., Jegelka, S., et al. (2018). Representation learning on graphs with jumping knowledge networks. *arXiv:1806.03536v2*. doi: 10.48550/arXiv.1806.03536

Yang, J., Liu, Z., Xiao, S., Li, C., Lian, D., Agrawal, S., et al. (2021). GraphFormers: GNN-nested transformers for representation learning on textual graph. *arXiv:2105.02605v3*. doi: 10.48550/arXiv.2105.02605

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., et al. (2021). Do transformers really perform bad for graph representation? *ArXiv abs/2106.05234*. doi: 10.48550/arXiv.2106.05234

Yu, W. H., Cuervo, A. M., Kumar, A., Peterhoff, C. M., Schmidt, S. D., Lee, J. H., et al. (2005). Macroautophag--a novel β-amyloid peptide-generating pathway activated in Alzheimer's disease. *J. Cell Biol.* 171, 87–98. doi: 10.1083/jcb.200505082

Zhang, M., Cui, Z., Neumann, M., and Chen, Y. An end-to-end deep learning architecture for graph classification. In 32nd AAAI Conference on artificial intelligence, AAAI 2018, (2018).

Zhang, P., Qin, W., Wang, D., Liu, B., Zhang, Y., Jiang, T., et al. (2015). Impacts of PICALM and CLU variants associated with Alzheimer's disease on the functional connectivity of the hippocampus in healthy young adults. *Brain Struct. Funct.* 220, 1463–1475. doi: 10.1007/s00429-014-0738-4

Zhang, Z., Xu, X., Tian, W., Jiang, R., Lu, Y., Sun, Q., et al. (2020). ARRB1 inhibits non-alcoholic steatohepatitis progression by promoting GDF15 maturation. *J. Hepatol.* 72, 976–989. doi: 10.1016/j.jhep.2019.12.004

Zhang, J., Zhang, H., Sun, L., and Xia, C. G.-B. (2020). Only attention is needed for learning graph representations. *ArXiv 2001.05140v2*. doi: 10.48550/arXiv.2001.05140

Zhu, Y., Hou, H., Rezai-Zadeh, K., Giunta, B., Ruscin, A., Gemma, C., et al. (2011). CD45 deficiency drives amyloid-β peptide oligomers and neuronal loss in Alzheimer's disease mice. *J. Neurosci.* 31, 1355–1365. doi: 10.1523/JNEUROSCI.3268-10.2011