Cell Function Graphics: TOGGLE delineates fate and function within individual cell types via single-cell transcriptomics

Junpeng Chen[1,2,16+*], Zhouweiyu Chen[3+], Tianda Sun[5+], Eric Jiang[4+], Kaiqing Liu[1,2+], Yibing Nong[1], Tao Yuan[8], Pawan Daga[9], Charles C. Dai[10,11], Yexing Yan[13], Jinwen Ge[6,7], Haihui Wu[14], Tong Yang[6], Shanshan Wang[6], Zixiang Su[15], Tian Song[17], Ahmed Abdelbsset-Ismail[1,2], You Li[4]*, Changping Li[12]*, Richa A. Singhal[16]*, Kailin Yang[6,7]*, Lu Cai[2*], Alex P. Carll[1,16]*

1 Address: Department of Physiology, University of Louisville School of Medicine, Louisville, Kentucky,40204, USA

2 Address: Pediatric Research Institute, Department of Pediatrics, University of Louisville School of Medicine, Louisville, Kentucky 40202, USA

3 Address: Xiangtan University, Xiangtan 411105, China

4 Address: Kebei Hospital of Shaoxing, Shaoxing City 312081, China

5 Address: Department of Computer Science, University of York, Heslington, York YO10 5DD, United Kingdom

6 Address: Key Laboratory of Hunan Province for Integrated Traditional Chinese and Western Medicine on Prevention and Treatment of Cardio-Cerebral Diseases, School of Integrated Chinese and Western Medicine, Hunan University of Chinese Medicine, Changsha, China

7 Address: Hunan Academy of Chinese Medicine, Changsha, China

8 Address: College of Mechanical Engineering and Automation, Huaqiao University, Xianmen, China, yuan__t@163.com

9 Address: Department of Internal Medicine, University of Louisville, Kentucky, USA

10 Address: Department of Oral and Maxillofacial Surgery, School of Dentistry, University of Maryland, Baltimore, Maryland 21201, USA.

11 Address: Fischell Department of Bioengineering, A. James Clark School of Engineering, University of Maryland, College Park, Maryland 20742, USA

12 Address: School of mechanical engineering and automation, Fuyao University of Science and Technology, Fuzhou, China

13 Address: Department of Psychology, Daqing Hospital of Traditional Chinese Medicine, Daqing, China

14 Address: Pinghu Hospital of Traditional Chinese Medicine, Pinghu, Zhejiang Province, China

15 Address: University of Minnesota, College of Biological Sciences, St. Paul, MN 55108, USA, suzixiang.mitchell@gmail.com

16 Address: Center for Cardiometabolic Science, Division of Environmental Medicine, Christina Lee Brown Envirome Insttitute, University of Louisville, Louisville, KY 40202, United States.

17 Address: Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing, China

+ Junpeng Chen (drchenjunpeng@hotmail.com), Eric Jiang (jyjiang@link.cuhk.edu.hk), Zhouweiyu Chen (yahoo_con_cn@163.com) and Kaiqing Liu are co-first authors who have contributed an indispensable part to this article and should be considered to have made equal contributions. Their names are listed without any specific order.

*Corresponding author should be: Junpeng Chen (drchenjunpeng@hotmail.com), Changping Li (15073200235@163.com), You Li (liyou@cjgs23.wecom.work), Kailin Yang (yklab2014@hotmail.com) , Richa A. Singhal (richa.singhal@louisville.edu), Lu Cai (lu.cai@louisville.edu) and Alex P. Carll (Alex.Carll@louisville.edu). Alex P. Carll is the project and Lab leader.

Abstract

Functional RNA plays a crucial role in regulating cellular processes throughout the life cycle of a cell. Identifying functional changes at each stage—from inception to development to maturation, functional execution, and eventual death or pathological transformation—often requires systematic comparisons of functional expression across cell populations. However, because cells of the same type often exhibit similar gene expression patterns regardless of function or fate, it is challenging to distinguish the stages of cellular fate or functional states within the same cell type, which also limits our understanding of cellular memory. Cells of the same type that share structural and gene expression similarities but originate from different regions and perform slightly distinct functions often retain unique epigenetic memory signatures. Although RNA serves as a key executor of fundamental cellular functions, its high expression similarity among cells of the same type limits its ability to distinguish functional heterogeneity. To overcome this challenge, we developed TOGGLE, utilizing higher-resolution analytical methods to uncover functional diversity at the cellular level. Then we based on TOGGLE developed an innovative Graph Diffusion Functional Map, which can significantly reduce noise, thereby more clearly displaying the functional grouping of RNA and enabling the capture of more subtle functional differences in high-dimensional data. Ultimately, this method effectively removes the influence of baseline functions from classification criteria and identifies key trajectories of cell fate determination.

1 Introduction

Functional RNA plays a crucial role in regulating cellular processes throughout the life cycle of a cell. Identifying functional changes at each stage—from inception to development to maturation, functional execution, and eventual death or pathological transformation—often requires systematic comparisons of gene expression across cell populations. However, because cells of the same type often exhibit similar gene expression patterns regardless of function or fate, it is challenging to distinguish "initiation-process-termination" phases of cellular fate or functional states within the same cell type. This limitation is particularly relevant in the rapidly evolving field of cellular aging research. For example, the process of isolating cell samples undergoing a specific type of death is particularly challenging because different modes of programmed cell death overlap in their marker genes and complex signaling pathways. In instances of ischemic stroke, current analytical tools for single-cell RNA sequencing (scRNA-seq) can discern neuronal subtypes (e.g., glutamatergic neurons), but they lack the resolution to distinguish neurons at different stages of programmed cell death.[1] This limitation

constrains the identification of early intervention targets and hinders a deeper understanding of the underlying mechanisms.

The inability to distinguish within transcriptomic data the fates and functions of common cell types also limits our understanding of cellular memory. Cells of the same type that share structural and gene expression similarities but originate from different regions and perform slightly distinct functions often retain unique epigenetic memory signatures. For instance, research has recently revealed that neural stem cells (NSCs) undergo significant methylation changes before and after activation.[2] Although their classification remains unchanged, their functional properties can shift substantially via these epigenetic modifications. Since other methods have difficulty distinguishing different specific gene regulation patterns in the same type of cells, this study used DNA methylation analysis to distinguish the activation state of NSCs, thereby comparing changes in NSC function after changes in methylation modifications[2]. However, this approach has significant limitations. Single-cell whole-genome DNA methylation data typically covers only about 5% of the genome, resulting in extremely sparse data matrices with over 95% missing values. Developing an algorithm capable of effectively distinguishing the functional activation states of NSCs from scRNA-seq and single-cell epigenetics sequence data would allow for a more direct and comprehensive analysis of functional gene expression. We present an algorithm capable of performing multimodal analysis on scRNA-seq data and single-cell epigenomic sequencing data, enabling the identification of NSC functional activation states without the need for serial measurements in Section 2.7.

DNA transcription is a fundamental determinant of cellular functions such as energy metabolism, material transport, and signal transduction, all of which are essential for cell survival. As a result, cells of the same type exhibit highly similar RNA expression patterns. First, the conserved nature of metabolic and signaling networks ensures the stability of RNA expression levels. Even when individual cells display slight functional differences, their overall RNA expression framework remains largely unchanged. Second, intracellular regulatory mechanisms—including feedback control, epigenetic modifications, and RNA degradation—strictly maintain RNA homeostasis, preventing significant fluctuations in gene expression patterns. Moreover, many functional differences among cells are not solely determined by RNA levels but are influenced by factors such as post-translational modifications, protein interactions, and subcellular localization. These regulatory layers further reduce capacity to distinguish functional states within the same cell type simply from RNA. Additionally, technical limitations of RNA sequencing—such as noise, the difficulty of detecting low-abundance genes, and the homogeneity of co-expression patterns—make it challenging to rely solely on RNA expression for functional differentiation. Thus, while RNA serves as a key executor of fundamental cellular functions,

its high expression similarity among cells of the same type limits its ability to distinguish functional heterogeneity. To overcome this challenge, higher-resolution analytical methods are needed to uncover functional diversity at the cellular level.

To address these challenges, we introduce a novel cell function mapping approach based on functional derivation relationships. Cellular functional progression follows a tree-like structure, where RNA transcription exhibits systematic hierarchical changes. At its core, fundamental biological processes such as energy metabolism and material transport serve as the "trunk," from which cells diverge along different pathways to form specialized functional branches. We employ graph diffusion to construct a hierarchical representation of RNA expression. By generating a weighted similarity metric and leveraging its progressive nature, this method captures the layered structure of gene regulatory networks. When combined with self-supervised learning, the model can automatically identify classification boundaries at functional transition points, enabling precise delineation of cell clusters. The high-similarity regions generated through graph diffusion reveal specialized functions, overcoming the limitations of traditional clustering methods. This approach provides a high-resolution functional hierarchy, allowing for a refined distinction between broad classifications and functionally distinct cellular subtypes.

Ultimately, this method effectively removes the influence of baseline functions from classification criteria and identifies key trajectories of cell fate determination. We define this new topic as "Fate Tracing", as a question for tracking the history of cell functions/programmed process. By dynamically reconstructing RNA expression hierarchies, it serves as a powerful computational tool for single-cell transcriptomics and cellular engineering. In Section 2.6, we develop a novel RNA-based graph model, while the Methods section outlines four key data challenges. Readers interested in the computational details can refer to those sections.

2 Results

In our study of cell fate and functional transitions, we employ a multi-layered approach that integrates experimental and computational analyses to comprehensively assess cellular behavior under different conditions. First, we evaluate performance using developmentally labeled cells (Section 2.1), tracking their responses in specific growth environments to analyze their fate trajectories and functional changes. Next, we test genetically modified cells derived in vitro (Section 2.2), using gene editing to manipulate key regulatory factors and assess their impact on differentiation and functional plasticity. This allows us to compare the stability of naturally developing cells with that of artificially modified ones. At the gene expression level, we apply an unlabeled approach (Section 2.3) to identify functional RNA pathways, extending the analysis to RNA network-based signaling pathway discovery (Section

2.4). This enables the identification of key regulatory factors involved in cellular state transitions and broadens the scope to include intercellular communication systems.

After validating the effectiveness of our algorithm, we constructed a trajectory of cellular functional progression (Section 2.5) and integrated biological validation methods to confirm the identified pathways, ultimately uncovering functional targets. First, we identified cell clusters associated with programmed cell death (Section 2.5.1), demonstrating that even without predefined labels, gene expression patterns can reveal cells undergoing programmed death. We then mapped the complete trajectory of dying cells (Section 2.5.2), illustrating how different cell types progressively enter the death process. Building on this, we identified cells approaching ferroptosis (Section 2.5.3) by analyzing mixed signaling pathways, allowing us to pinpoint key molecular events preceding ferroptosis. To validate these findings, we conducted experimental verification of ferroptosis-related genes (Section 2.6). Through tissue analysis, we observed pathological changes (Section 2.6.1) and ultrastructural alterations (Section 2.6.2) in brain tissue under different conditions, providing morphological evidence for cell death mechanisms. Given the critical role of ferroptosis- and apoptosis-related pathways—including the GPX4/ACSL4 axis, FSP1/CoQ10 pathway, iron metabolism, TGF-β/Smad pathway, and NF-κB and Bcl-2/Bax pathways—in ischemic stroke-induced programmed cell death[3–5], we examined the expression levels of key genes within these pathways. Specifically, we used immunohistochemistry (Section 2.6.3) to assess the protein expression of Acsl4, Gpx4, and Fsp1. Additionally, we performed immunofluorescence staining (Section 2.6.4) to measure the expression of Bax, IkBa, Casp3, p-Smad, p65 (Rela), Tfrc, Trf, and Sirt1, providing deeper insights into the roles of these signaling pathways in the cell death process.

In Section 2.5, we enhance the resolution of scRNA-seq data analysis using TOGGLE. Ferroptosis research is closely linked to clinical medicine, and after performance validation through the experiments in Sections 2.1 to 2.5, we provide a detailed explanation in Section 2.5 on how TOGGLE can be utilized for identifying potential drug pathways. By classifying different stages of cell death, TOGGLE enables precise tracking of the programmed progression of cells, allowing for the independent development of targeted therapies for early, progressive, and terminal stages based on medical needs. Furthermore, in conjunction with the findings in Section 2.4, TOGGLE facilitates intercellular communication analysis to uncover how the microenvironment influences cell behavior, opening new possibilities for stem cell therapy in treating degenerative diseases, tissue damage, and cancer.

Distinguishing the activation states of NSCs is particularly challenging due to the highly similar RNA transcription processes of neural cells and the significant effort required to replicate complex in

vivo conditions in vitro. Restoring neural function is one of the most difficult steps in organ regeneration, as neurons have almost no regenerative capacity, making it extremely challenging to assess their functionality outside the body. Before transplanting regenerated organs into patients, extensive and costly environmental simulations are typically required to verify neuronal activity and ensure functional integrity. We successfully differentiated active and quiescent neural stem cells, revealing that DNA methylation memory is a key factor driving cells into a metabolically active state. This discovery enables us to determine the functional state of neural cells by directly capturing their methylation memory, eliminating the need for additional culture steps to allow full RNA function expression. This approach not only streamlines the assessment of neural cell function but also significantly reduces time and costs..

2.1 Performance Testing Using Developmental Cells

We employed an unsupervised algorithm to compare various methods that require predefined cell types, recognizing that cellular function and lineage transitions are continuous and dynamic processes. Ideally, if a comprehensive and representative set of labeled samples covering the entire cell fate spectrum were available, classifying cell states would be straightforward without the need to distinguish function from lineage. However, existing cell-tracing techniques are not only costly but also ineffective for mature, non-dividing cells. Additionally, due to inherent systematic biases in biological experiments, integrating scRNA-seq profiles remains challenging, even when addressing the same problem repeatedly. This limitation makes supervised learning methods impractical for real-world applications. Therefore, developing a tool that can accurately analyze cell fate trajectories without relying on predefined labels is essential. To validate our approach, we first tested the algorithm on developing progenitor cells, where clear identification labels were available, before applying it to mature cells for biological discovery.

To assess the efficacy of TOGGLE and make comparisons with other methods, we applied it to publicly available datasets with known lineage. The first dataset traced the differentiation of mouse bone marrow hematopoietic progenitor cells (HPCs) in a controlled culture environment (GSE140802) over four timepoints[6]. We identified cells using the method in Weinreb et al.'s work[6] and focused our analysis on three cell types: progenitor cells, monocytes, and neutrophils: progenitor cells, monocytes, and neutrophils. Progenitor cells exhibit a dual fate bias, as defined in the original Cospar research paper[7], where a portion differentiates into monocytes (referred to as monocyte progenitors), while another portion differentiates into neutrophils (referred to as neutrophil progenitors)[8]. Our objective is to trace the fate bias of progenitor cells, specifically to identify monocyte progenitors and neutrophil progenitors within the progenitor cell population.

TOGGLE is an unsupervised method. TOGGLE first generates a similarity matrix described by graph distances (Section 4.2). Cells that are relatively close are considered to have the same fate preferences and are clustered into a unified group in the same iteration round. The algorithm segments the distance matrix through binarization (Section 4.3), initially obtaining a boundary, then iteratively generating a second boundary based on the first (Section 4.4). Each classification results in a positively correlated group and a negatively correlated group, with the boundary between them forming a dividing line. Within the positively correlated group, a secondary positively correlated group and a secondary negatively correlated group are identified. Similarly, within the negatively correlated group, a secondary positively correlated group and a secondary negatively correlated group are identified. This process creates a total of two new boundaries at the intersections (Section 4.4). This process results in different groupings of cells, which are then sorted. After sorting, then deriving the required number of categorie according to the resolution set by number of target categories boundaries (Section 4.5). After completing the segmentation, samples within adjacent boundaries form a cluster. In this process, progenitor cells close to neutrophils are determined by TOGGLE to be neutrophil progenitors, while progenitor cells close to monocytes are determined to be monocyte progenitors (Figure 1G). (1) When analyzing descendants independently, by comparing with the true tracing labels in ground truth analysis, the accuracy of TOGGLE is found to be 79%, predict result as Figure 1H. Notably, the TOGGLE algorithm is implemented as a unsupervised method without prior training. In order to evaluate the performance of TOGGLE, we compared it to four widely used in silico algorithms: Cospar, Super OT, WOT, and GAN-Based OT. Cospar attained a rate of 65%, Super OT exhibited a slightly higher accuracy at 67%, WOT maintained a similar accuracy of 65%, and GAN-Based OT achieved 62%. (2) When both progenitor and descendant cells are selected for classification, TOGGLE achieves an accuracy of 82%. The accuracy of WOT is less than 50%, while Cospar cannot be used for this calculation.

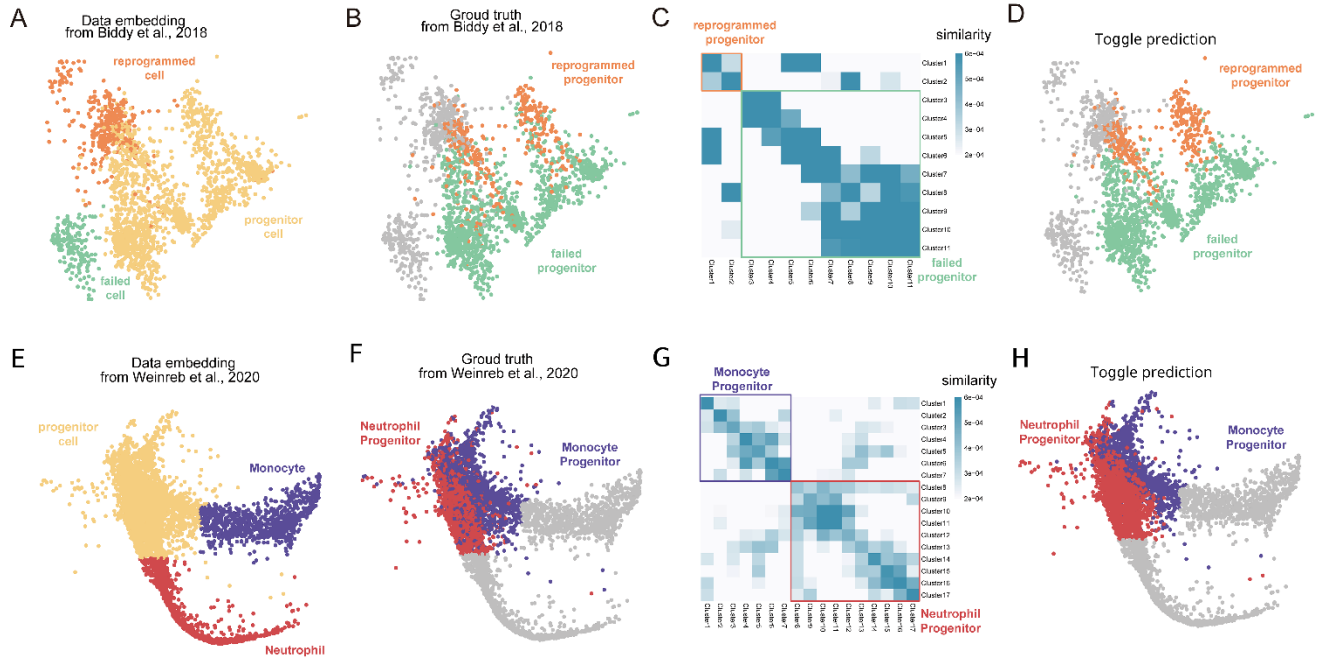

Figure 1. Tracing the lineage of fibroblast reprogramming[9] and hematopoietic cell[16] development. A: Data embedding of scRNA-seq data on experimentally-induced fibroblast reprogramming over six sampled timepoints in Biddy et al., 2018[9]; B: Ground truth of reprogrammed data; C: Similarity matrix of clustering results of reprogrammed data; D: TOGGLE's prediction result of reprogrammed data; E: Data embedding of neutrophil progenitor data from Weinreb et al. 2020[6]; F: Ground truth of neutrophil progenitor data; G: The similarity matrix of clustering results of neutrophil progenitor data; H: The TOGGLE's prediction result of neutrophil progenitor data.

2.2 Testing with Genetically Modified In Vitro Cells

We also validated the performance of TOGGLE on a dataset from mouse embryonic fibroblasts sampled at eight timepoints across transcription factor (TF) direct lineage reprogramming to induced endoderm progenitor cells (GSE99915)[9]. In this experiment, cells initially expanded, but many were lost due to aging or cell death. By day 28, 30% of the cells were successfully reprogrammed[9]. Building on the work of Biddy et al.[9], we identified cells using genes such as *Spint2, Apoa1, Peg3, and Col1a2*, which are strongly correlated with successful or failed reprogramming. For further analysis, we focused on progenitor cells (fibroblasts), reprogrammed cells, and failed cells. Induced progenitor cells display a dual fate bias, wherein a fraction differentiates into reprogrammed cells, referred to as reprogrammed progenitors, while another fraction differentiates into failed-reprogrammed cells, referred to as failed progenitors[9]. Here, we investigated the fate bias of fibroblasts undergoing transcription factor-induced reprogramming to progenitors, specifically by identifying the successfully reprogrammed progenitors and failed progenitors within the progenitor cell population.

We employed TOGGLE to reconstruct the complete tracing of the reprogramming process of fibroblasts to stem cells, methods same as section 2.1. After completing the segmentation, samples within adjacent boundaries form a cluster. In this process, progenitor cells close to Reprogrammed cells are identified by TOGGLE as Reprogrammed progenitor cells, while those close to Failed cells are identified as Failed progenitor cells (Figure 1C). (1)When analyzing descendants independently, by comparing with the true tracing labels in the ground truth, TOGGLE achieved an accuracy of 91.32%, predict result as Figure 1D. TOGGLE is a unsupervised algorithm, and we benchmarked it against the two most common in silico algorithms, with Cospar's accuracy being 67%, and WOT's accuracy being 31%, showing a significant accuracy advantage. (2)When both progenitor and descendant cells are selected for classification, TOGGLE achieves an accuracy of 84%. The accuracy of WOT is less than 50%, while Cospar cannot be used for this calculation.

2.3 Label-Free Discovery of RNA Functional Pathways

TOGGLE was applied to analyze the effects of e-cigarette aerosol inhalation on mouse cardiac tissue using dataset GSE183614[10]. Taking the differentially expressed genes of adult mice exposed to e-cigarette smoke and fresh air as an example, they were divided into 6 groups (Section 4.5). Except for Group 2, Group 3, Group 4 and Group 5, the remaining groups did not show significant enrichment in biological processes or KEGG signaling pathways and require further investigation (Figure 2). The enrichment results of each gene group and the grouping results of differentially expressed genes for other intervention methods are shown in Supplementary Material 7. The results indicate that exposure to e-cigarette aerosol significantly altered RNA pathways in the hearts of mice. These changes are reflected in the grouping of differentially expressed genes and their enrichment in biological functions and signaling pathways. More pronounced immune and inflammatory responses (e.g., TNF signaling) and cell cycle-related changes among young mice compared to mature adults suggested that the impact of e-cigarette use is greater at younger ages.

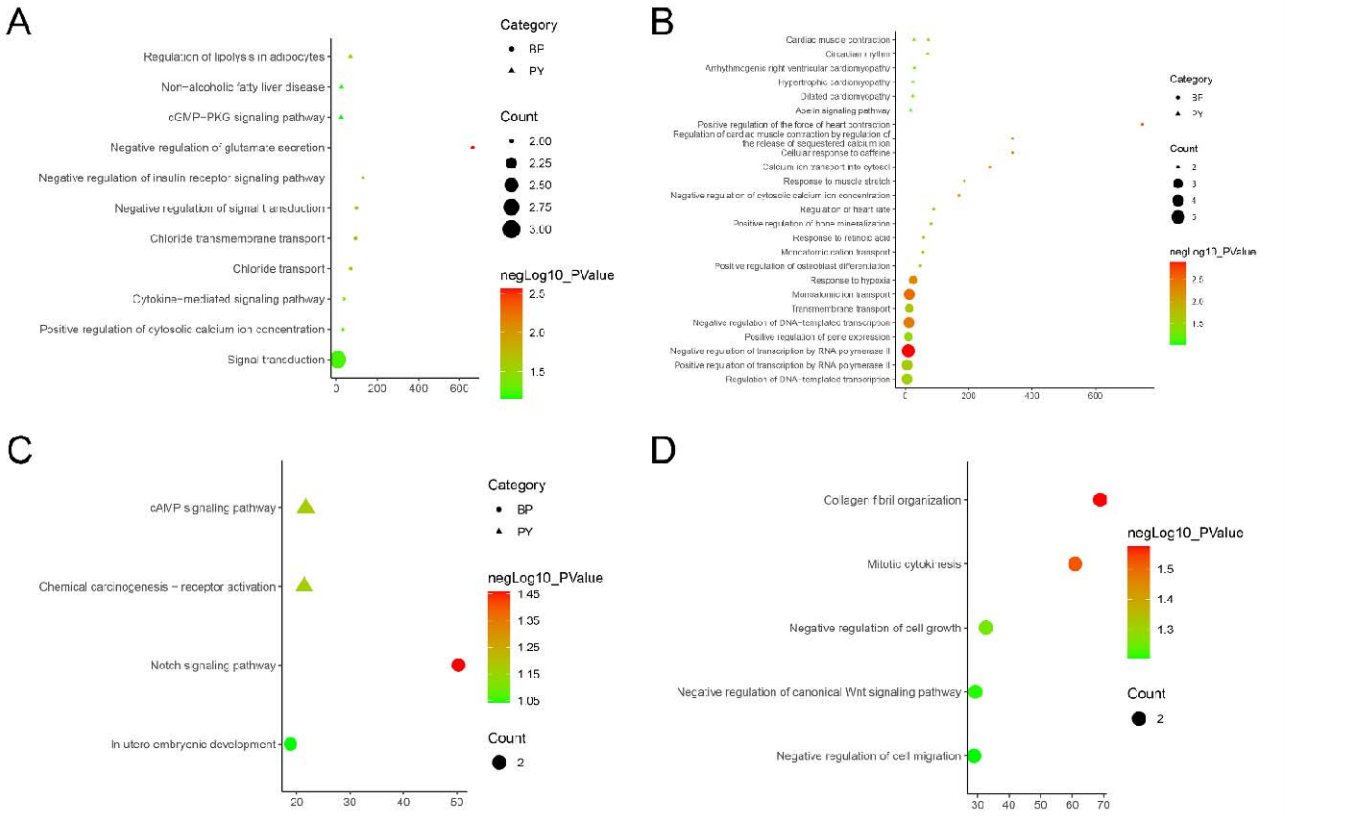

Figure 2. The Enrichment Results of Genes in Each Gene Group. A: Biological processes and pathways of Group 2; B: Biological processes and pathways of Group 3; C: Biological processes and pathways of Group 4; D: Biological processes of Group 5. (BP: biological processes; PY: signaling pathways. X-axis is fold enrichment)

2.4 Cellular functional mapping: Discovery of Signaling Pathways in the RNA Network

TOGGLE was applied to analyze single-nucleus RNA sequencing (snRNA-seq) and bulk RNA sequencing (bulk-RNA seq) data from a rat myocardial infarction (MI) model (dataset: GSE253768). After classifying the snRNA-seq data into various cell types, including fibroblasts, myofibroblasts, macrophages, T cells, and cardiomyocytes (Figure 3A), TOGGLE divided the fibroblasts and myofibroblasts into seven groups (Figure 3B). The MI group was predominantly represented by FibR1-G5, FibR1-G6, and FibR1-G7 (Figure 3C). The results of the inferred ligand-receptor interactions between these fibroblasts and T cells or macrophages (cell-cell communication) revealed differences in the strength and types of interactions, suggesting functional variations among these fibroblast subgroups (Figures 3D and 3E). Further analysis of the highly variable genes in the MI fibroblasts also identified eight distinct functional groups (Figure 3F). Additionally, TOGGLE was used to map bulk RNA-seq data onto the snRNA-seq dataset, revealing that the gene expression pattern of fibroblasts in the bulk RNA-seq MI group closely matched that of FibR2-G6 (highlighted in the red box in Figure 3G). This

finding suggests that the bulk RNA-seq fibroblasts predominantly exhibit functions similar to those of FibR2-G6, which are associated with cell migration, adhesion, angiogenesis, and ECM-receptor interactions (Figure 3H).

Thus, TOGGLE can effectively combine bulk RNA-seq data with snRNA-seq to contextualize each, and which cell types and specific phenotypes undergo nuclear transcriptional changes that most closely represent overall tissue-level changes in transcription.

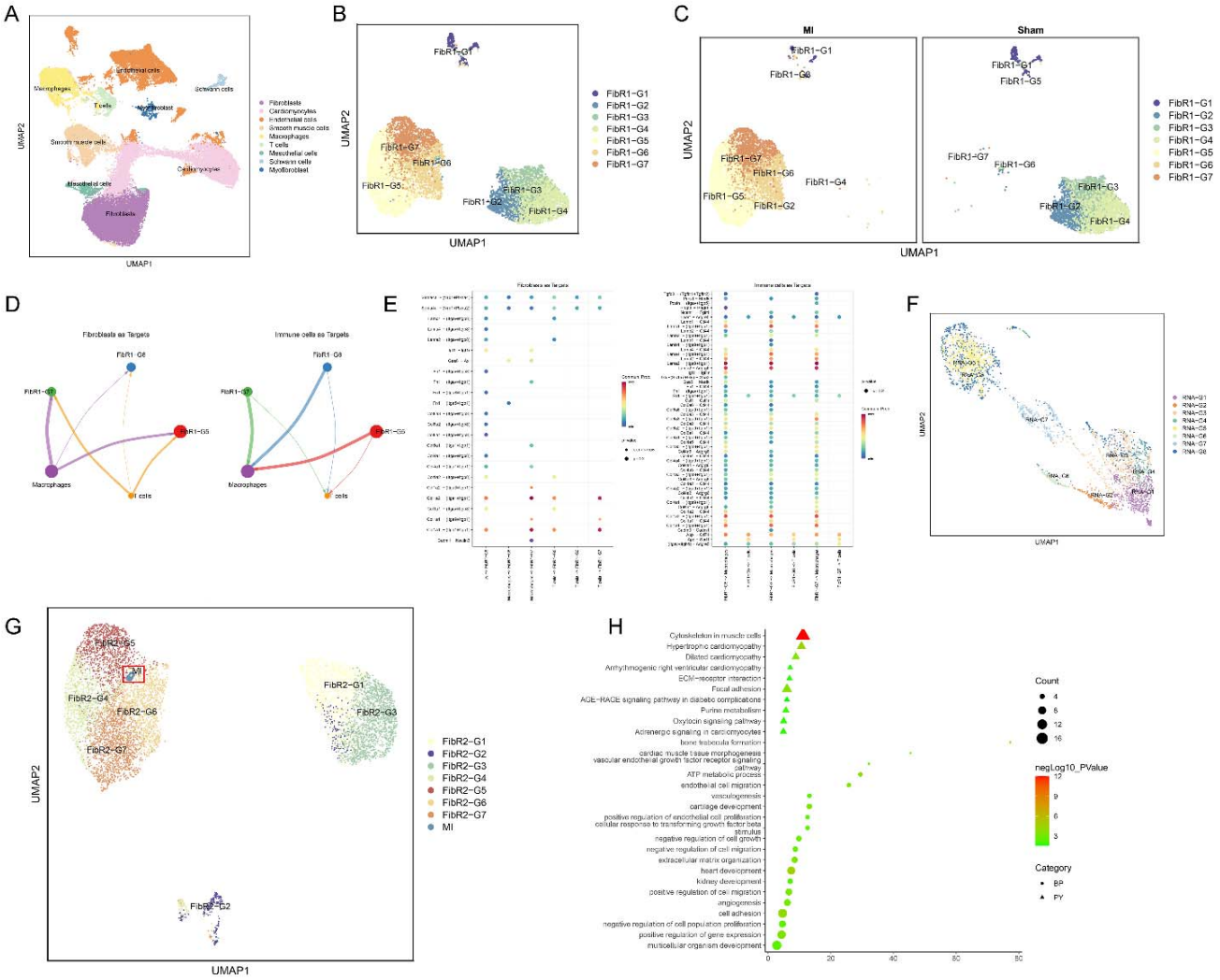

Figure 3 Analysis results of GSE253768 set. A: Cell types of snRNA-seq; B: Fibroblasts and myofibroblasts were divided into 7 groups after TOGGLE analysis; C: Distribution of 7 cell types in MI cohort and Sham cohort; D: Weight/strength diagram of cell communication between fibroblasts and immune cells; E: Bubble diagram of cell communication between fibroblasts and immune cells; F: RNA distribution of fibroblasts in the MI group; G: Mapping MI group bulk RNA-seq data onto UMAP of fibroblast snRNA-seq results; H: Enrichment results of FibR2-G6 (BP: biological processes; PY: signaling pathways. X-axis is fold enrichment).

2.5 Construction of Functional Change Trajectories and Target Identification

2.5.1 Identifying Programmed Cell Death-Related Cell Clusters Without True Labels

The above results confirm the reliability of TOGGLE for predicting cell fate. In cases of ischemic stroke, neurons often undergo programmed cell death, with ferroptosis playing a significant role in neuronal loss. However, ferroptosis frequently overlaps with other forms of programmed cell death, making it challenging for existing algorithms to effectively distinguish ferroptosis from other cell death mechanisms. Therefore, we applied TOGGLE to predict and analyze neuronal fate in a rat model of ischemic stroke, distinguishing between healthy and dying neurons to forecast programmed cell death, particularly ferroptosis. And we further verified the expression of apoptosis genes and ferroptosis genes in animal experiments. Ischemic stroke samples [middle cerebral artery occlusion (MCAO) cohort] and healthy samples (Sham cohort) from the GSE232429 dataset were included in the analysis[1]. The neuronal scRNA-seq data were then analyzed using TOGGLE and initially divided into five groups by dimension reduction using UMAP: Group R1-1, Group R1-2, Group R1-3, Group R1-4, and Group R1-5 (Figure 4A).

Cohorts were analyzed for differentially expressed genes, which we cross-referenced with programmed cell death-related genes[11] that we then counted in each comparison group (Figure 4D). Group R1-1 vs. Group R1-5 had fewer programmed cell death-related genes, while comparisons between Group R1-1 and other groups, as well as Group R1-5 and other groups, had a higher number of programmed cell death-related genes. This suggests that Group R1-1 and Group R1-5 exhibit similar states, potentially representing consecutive stages in a continuum. Additionally, the cell counts of Group R1-1 and Group R1-5 in the sham cohort were high (Figure 4B and 4C), suggesting that both groups consisted of cells that had not entered a programmed cell death state. Given that Group R1-5 was abundant in both MCAO cohort and sham cohort, we inferred that it represented healthy cells. Furthermore, Group R1-2 cells were more prevalent in the MCAO cohort compared to sham and had a significantly more differentially expressed programmed cell death-related genes compared to other cell groups, indicating that Group R1-2 consists of cells that have entered a programmed cell death state. Given the moderate number of differentially expressed programmed cell death-related genes among cells in Group R1-3 and Group R1-4, they are inferred to be in intermediate states.

Based on CellMarker, the cells of Group R1-3 were identified as Mtss2+Gpr37l1+ glutamatergic neurons, and that of Group R1-4 were identified as Atp2b2+Grin2a+ glutamatergic neurons. As shown in Figure 5B and 3C, the cell counts of Group R1-3 and Group R1-4 showed little difference between the MCAO and Sham cohorts, which suggests that these two groups of cells had not undergone

programmed cell death. Therefore, these groups were excluded from further analysis. The UMAP plots after excluding Groups R1-3 and R1-4 are shown in Figure 4E.

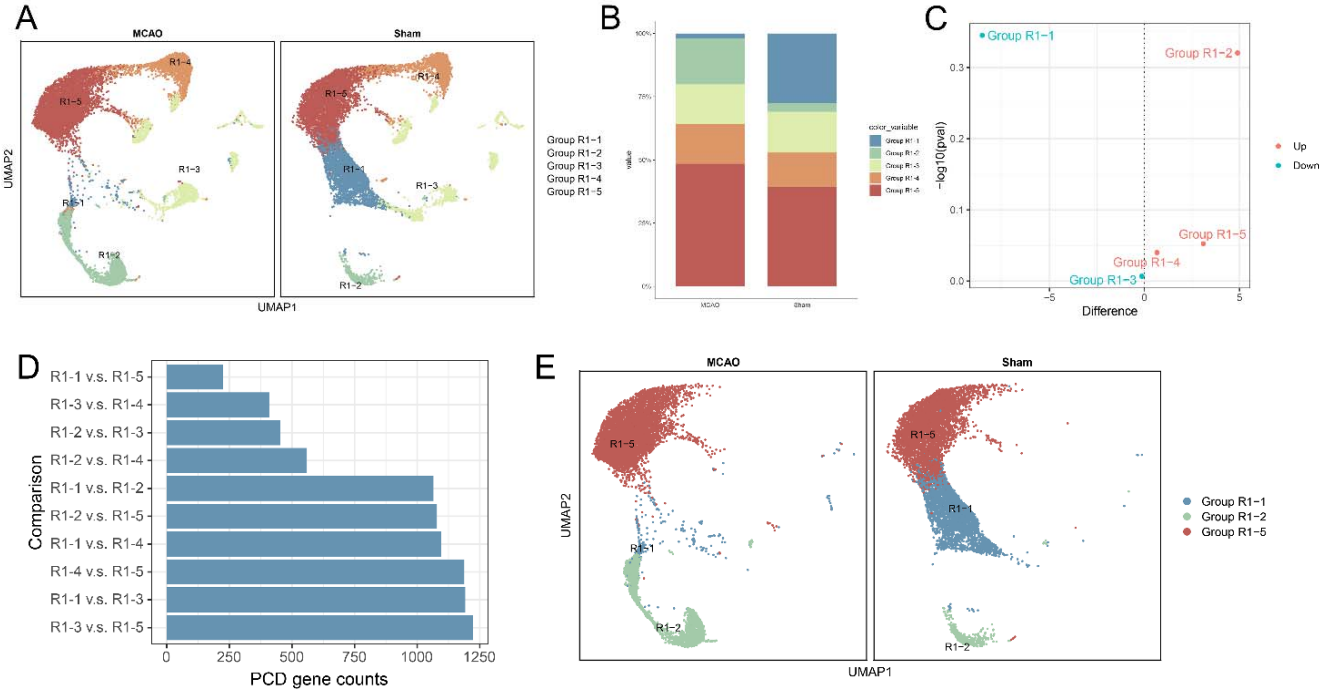

Figure 4. Preliminary Prediction of Neuronal Fate. A: The grouping results of the preliminary analysis; B: The proportion of each cell group in MCAO and Sham; C: Compared with Sham cohort, the changes in cell number in each cell group in MCAO cohort; D: Comparison of the number of differentially expressed programmed cell death genes between cell groups of all cohorts; E: UMAP graph without R1-3 and R1-4.

2.5.2 Construction of the Complete Programmed Death Trajectory

The data of Group R1-1, Group R1-2, and Group R1-5 of the preliminary analysis results were further analyzed to predict their programmed cell death trajectories. Monocle3 was utilized to analyze their pseudotime (Figure 5A), it can be observed that the loss of temporal information is significant because the time restoration is not synchronized with the trajectory scale. And based on this data, TOGGLE identified 9 subgroups (Figure 5B and 5C). Based on the number of programmed cell death-related genes and the cell count within each group (Figure 5D), the following can be inferred: (1) Groups R2-1 are closely related to Group R2-6. (2) Groups R2-3 are distant from Groups R2-5, R2-6, and R2-7; similarly, Groups R2-1 are distant from Groups R2-5 and R2-7, consistent with the pseudotime analysis. (3) Group R2-2 is closely related to Group R2-4; however, since the cell counts of Group R2-2 in the sham cohort is larger, it is considered potential noise (Figure 5E, 5F, 5G). R2-9 cells are likely representative of healthy cells. R2-8 cells are predominantly observed in the MCAO cohort, while R2-1, R2-5, R2-6, and R2-7 are more frequent in the Sham cohort, suggesting they may represent

intermediate states transitioning toward programmed cell death. In contrast, R2-2 and R2-3 cells show a marked increase in the MCAO cohort, indicating they are primarily composed of cells undergoing programmed cell death. The trajectory of programmed cell death may be Group R2-9→Group R2-8→Group R2-1→Groups R2-5 to R2-7→Group R2-2→Group R2-3.

In summary, we deduced that Group R2-9 is the most likely starting point and Group R2-3 the most plausible terminus. The FindMarkers function of Seurat was used to perform precise differential gene expression analysis. Group R2-1, Group R2-2, Group R2-3, Groups R2-5 to R2-7, and Group R2-8 were set as ident1, and Group R2-9 was set as ident2. The volcano plot of Group R2-9 v.s. Group R2-3 genes is shown in Figure 5H. The differentially expressed genes between these groups were imported into DAVID (ver. 2024q2, https://david.ncifcrf.gov) for Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathway enrichment analysis, so as to obtain the number of genes in representative programmed cell death pathways (such as apoptosis and ferroptosis) (Figure 5I and 5J). It was found that as the trajectory moved away from Group R2-9, the number of apoptosis- and ferroptosis-related genes expressed in each group gradually increased, which was consistent with the inferred time trajectory. The results of enrichment analysis are shown in Table S1 (see supplementary materials), and the differentially expressed genes are shown in Table S2.

Given the critical role of pathways related to ferroptosis and apopotosis (such as the GPX4/ACSL4 axis, FSP1/CoQ10 pathway, iron metabolism, and TGF-β/Smad pathway, NF-κB and Bcl-2/Bax pathways)[3–5] in programmed cell death caused by ischemic stroke, we focused on the expression of key genes in these pathways between Group R2-3 and Group R2-9 (Table S2). The expression of key genes in these pathways is as follows: (1) The expression patterns of most genes are consistent with those observed during apoptosis and ferroptosis: Gpx4 (Log2FC=-0.60, FDR<0.05) , Acsl4 (Log2FC=1.39, FDR<0.05), Trf (Log2FC=6.98, FDR<0.05), Tfrc (also known as TFR, Log2FC=3.38, FDR<0.05), Smad1 (Log2FC=1.16, FDR<0.05), *Smad5* (Log2FC=1.70, FDR<0.05), *Rela* (also known as *p65*, Log2FC=1.08, FDR<0.05), *Bax* (Log2FC=1.46, FDR<0.05), and *Nfkbia* (also known as *IkBA*, Log2FC=4.00, FDR<0.05). (2) However, several apoptosis- and ferroptosis-related genes showed expressions inconsistent with previous studies, including *Aifm2* (also known as *FSP1*, Log2FC=0.66, FDR<0.05), *Sirt1* (Log2FC=2.56, FDR<0.05), and *Casp3* (Log2FC=-0.40, FDR<0.05). This may be because cells in the early stages of programmed cell death are trying to "self-rescue." In addition, some positively expressed ferroptosis-driving genes were found in the differentially expressed genes. The ferroptosis-driving genes refer to FerrDb V2 (http://www.zhounan.org/ferrdb/current/)[12] (Table S2). The positions of these genes within relevant pathways are illustrated in Figure 6.

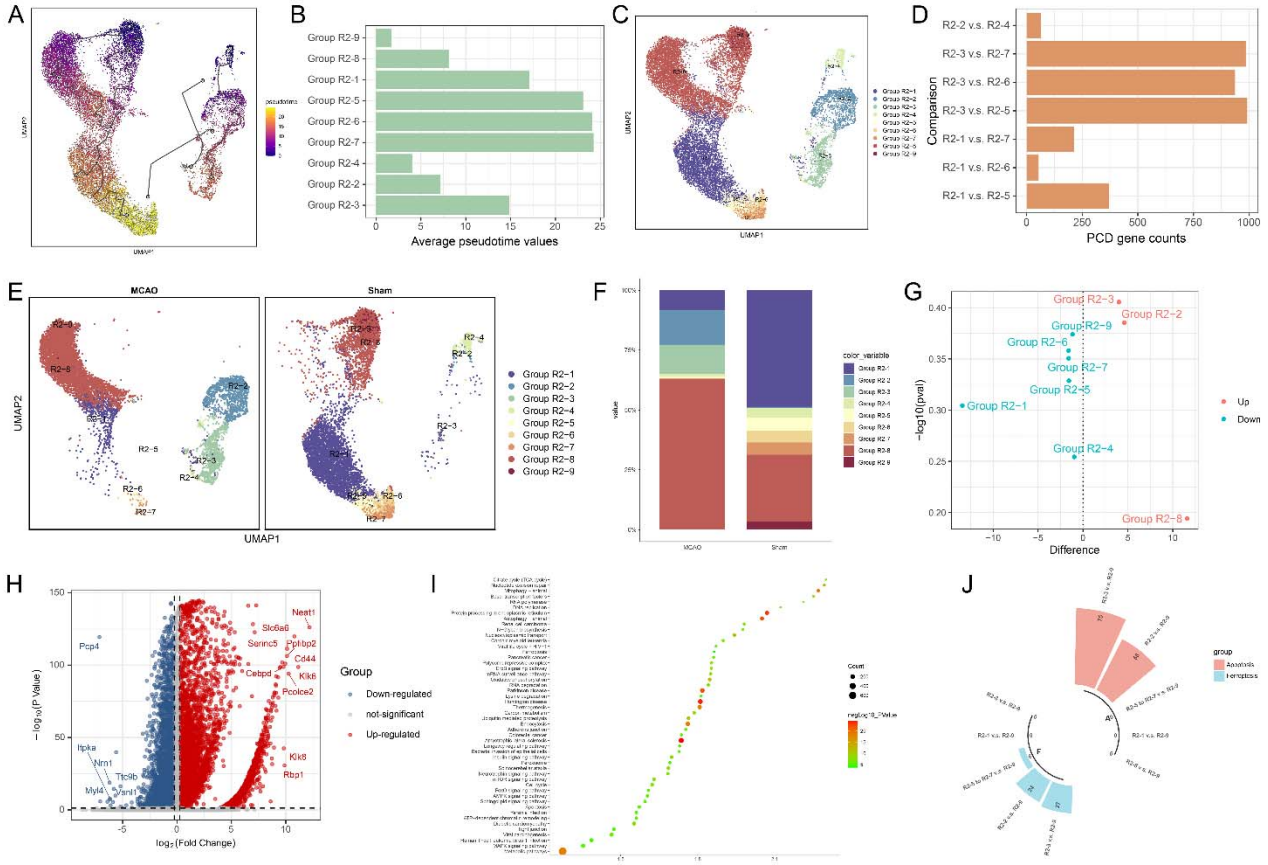

Figure 5. Programmed Cell Death Trajectories Prediction. A: Pseudo-time trajectory of cells; B: The average pseudotime values of groups; C: Cells in previous groups R1-1, R1-2, R1-5 further analyzed by TOGGLE that divided into 9 groups; D: The number of programmed cell death genes in each cell group; E: Distribution of cell groups in Sham and MCAO; F: The proportion of those 9 cell groups in MCAO and Sham; G: Compared with Sham, the changes in cell number in those 9 cell groups in MCAO; H: Volcano plot of Group R2-3 v.s. Group R2-9 genes; I: Top 50 pathways of Group R2-3 v.s. Group R2-9 genes; J: The number of genes enriched in apoptosis and ferroptosis. Pink genes were selected for experimental verification in follow-up studies (Figure 9 and Figure 10). PCD: programmed cell death.

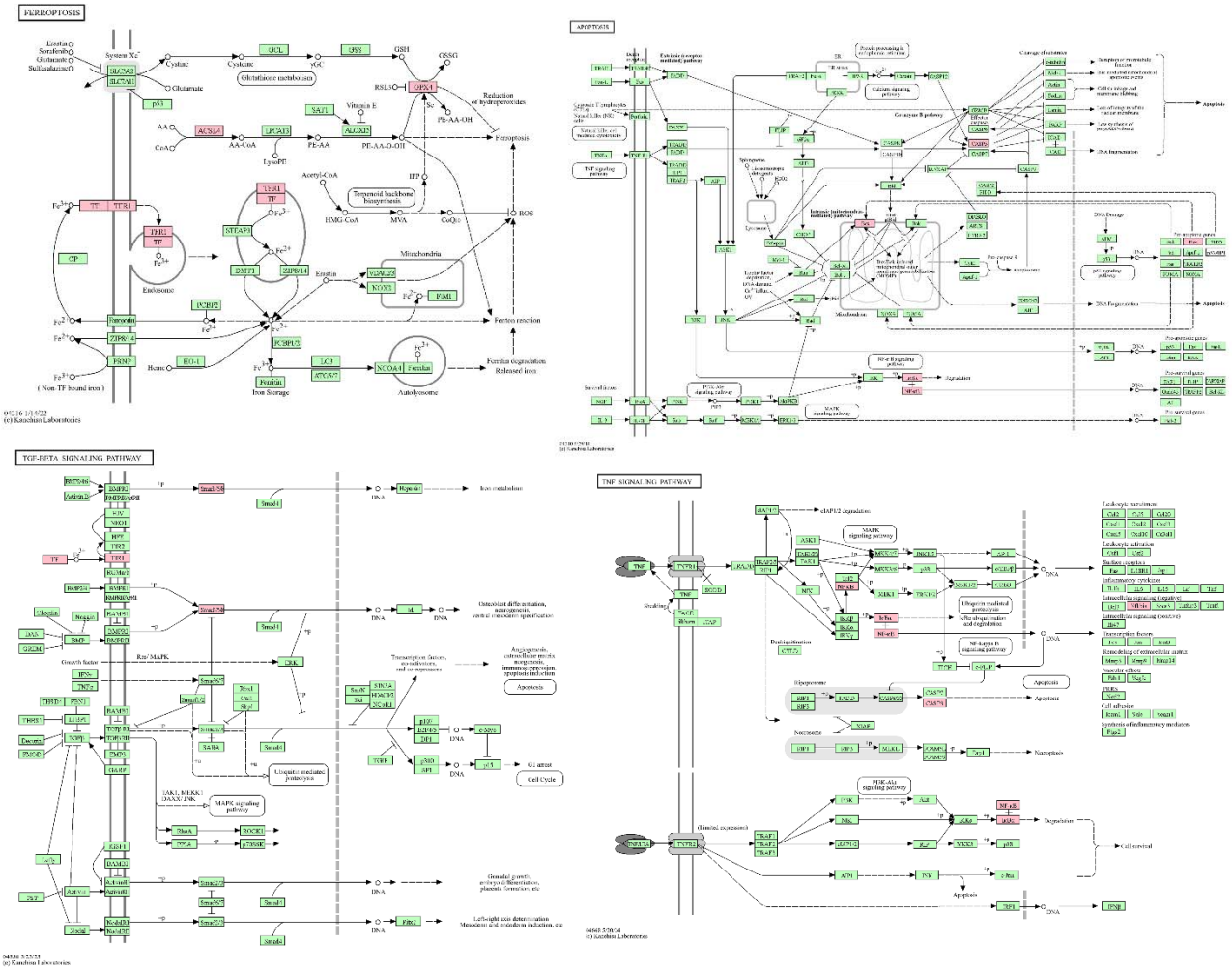

Figure 6. Signaling pathways related to ferroptosis and apoptosis. (modified from KEGG pathway mmu04216, mmu04210, mmu04350 and mmu04668[13])

2.5.3 Identification of Cells on the Brink of Ferroptosis from Confounding Pathways

To further analyze ferroptosis-dominant cell populations, TOGGLE was applied to Groups R2-2 and R2-3, and five groups (Figure 7A). Using the FindAllMarkers function of Seurat, genes uniquely and highly expressed in each group were identified with selection criteria of Log2FC ≥ 0.25 and FDR < 0.05. Ferroptosis driver genes were sourced from the FerrDb database (Ver. 2, http://www.zhounan.org/ferrdb/current/). Among the differentially expressed genes in Groups R3-4 and R3-5, ferroptosis driver and other programmed cell death genes were most prevalent. Furthermore, these groups displayed a higher number of highly expressed ferroptosis driver genes (Log2FC ≥ 2 and FDR < 0.05) compared to other groups, with over 70% of cells in Groups R3-4 and R3-5 showing ferroptosis markers relative to healthy cells (Figure 7B, 7C and 7D). The enrichment results showed that the

pathways enriched by differentially expressed genes included ferroptosis (Figure 7E). These findings support identifying Groups R3-4 and R3-5 as ferroptosis-dominant cell populations.

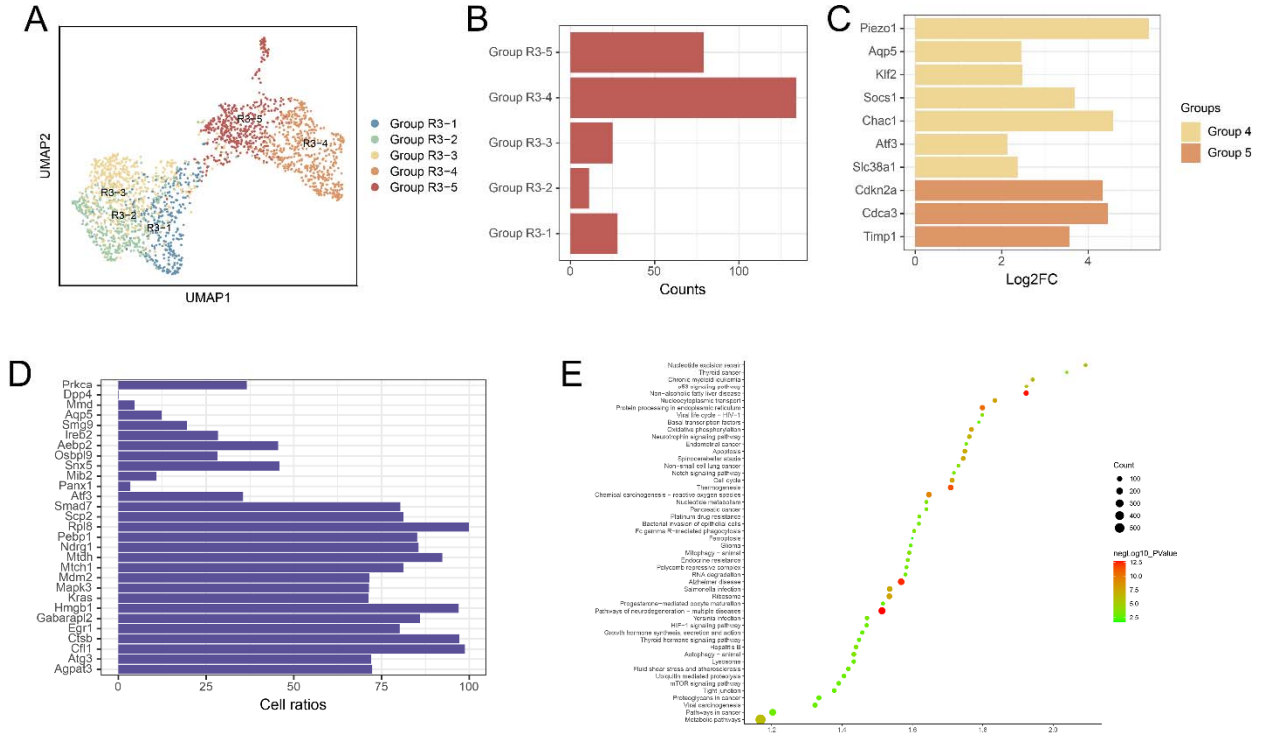

Figure 7. Ferroptotic Neurons Prediction. A: Cells in previous Groups R2-2 and R2-3 further analyzed by TOGGLE and divided into 5 groups; B: The proportion of ferroptosis-driving genes in each cell group; C: Log2FC values of highly expressed ferroptosis-driving genes; D: The proportion of cells expressing different ferroptosis-driving genes; E: Top 50 pathways of Group R3-4 and Group R3-5 genes

2.6 Verification of apoptosis- and ferroptosis-related genes

Our previous research demonstrated increased iron levels and neuronal iron accumulation in the infarcted brain tissue of the MCAO model. Deferiprone treatment in MCAO model effectively reduces iron content and diminishes iron accumulation[14]. Based on these findings, we used the MCAO model ± deferiprone to validate the predicted results of neuronal ferroptosis by TOGGLE.

2.6.1 Pathological changes of brain tissue

In the Sham operation group, neurons exhibited normal morphology, with orderly arrangement, clear nuclear membranes, uniformly stained cytoplasm, and no evident pathological changes. In contrast, the MCAO group showed significant structural damage in the ischemic cortical area, with disorganized neuronal arrangement, widened intercellular spaces, and signs of cell swelling, rupture, and nuclear condensation, leading to vacuole formation. In the MCAO+deferiprone group, neuronal structural

damage was mild, with minimal cell membrane rupture, slight nuclear condensation, and a significant reduction in vacuolar degeneration (Figure 8).

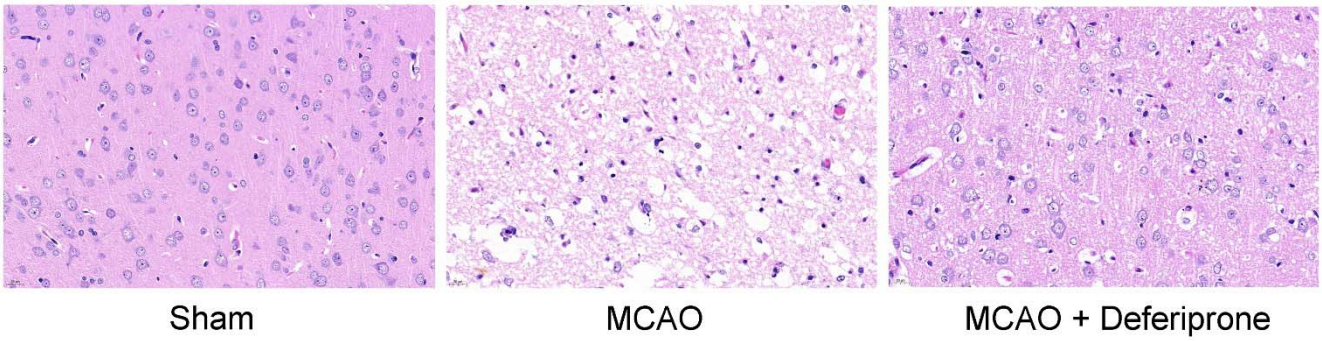

Figure 8 Pathological changes of brain tissue. HE staining, 40× magnification

2.6.2 Ultrastructural changes of brain tissue

In the Sham group, mitochondria appeared structurally normal with clear features and intact membranes. In contrast, the MCAO group exhibited mitochondrial atrophy, reduced cristae, and increased membrane density, consistent with the morphological characteristics of ferroptosis. The MCAO+deferiprone group showed significant improvement in mitochondrial integrity compared to the MCAO group, with reduced damage (Figure 9).

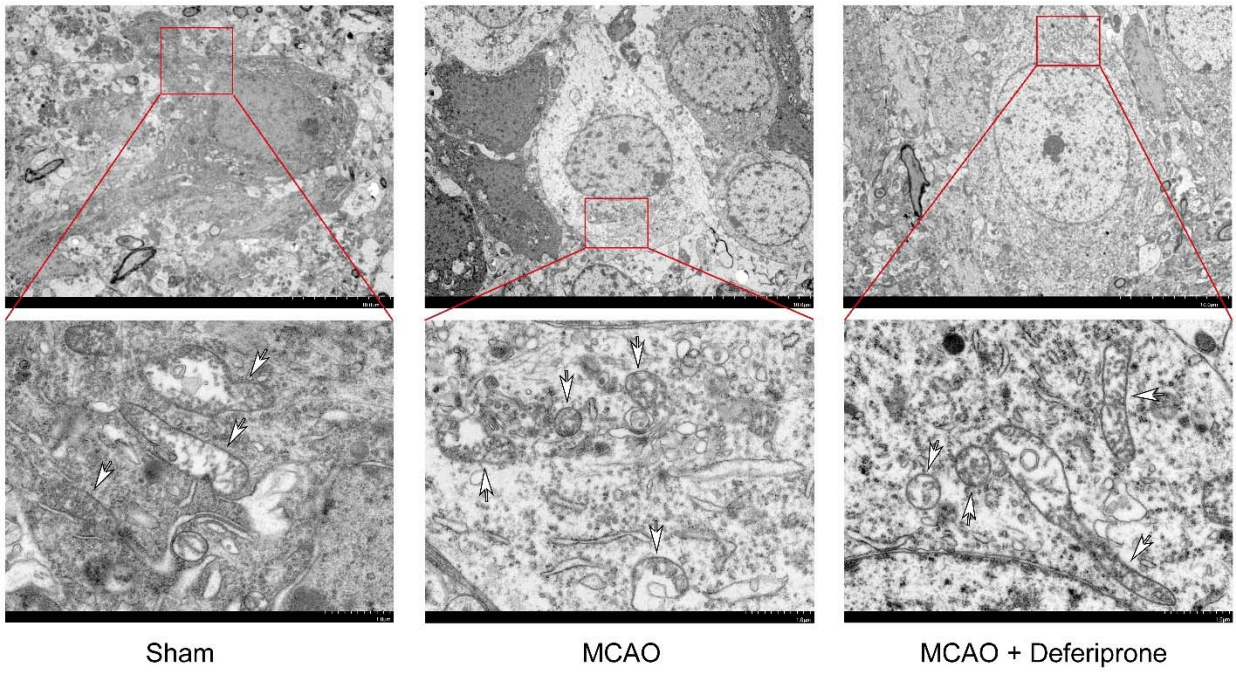

Figure 9 Ultrastructural changes of brain tissue. Transmission electron microscope. The upper images are at 1500× magnification, and the lower images are at 6000× magnification. Mitochondria are marked by white arrows.

2.6.3 Expression levels of *Acsl4, Gpx4* and *Fsp1* proteins detected by immunohistochemistry

Compared to the Sham operation group, the MCAO group showed an increase in Acsl4 levels ($P<0.05$) and a decrease in Gpx4 levels ($P<0.05$). These findings are consistent with the differential gene expression results from single-cell sequencing. Meanwhile, compared to the Sham operation group, the MCAO group showed an decrease in Fsp1 levels ($P<0.05$). Compared with the MCAO group, the MCAO+deferiprone group showed a decrease in Acsl4 levels and Fsp1 levels ($P<0.05$), and a increase in Gpx4 levels ($P<0.05$). (Figure 10).

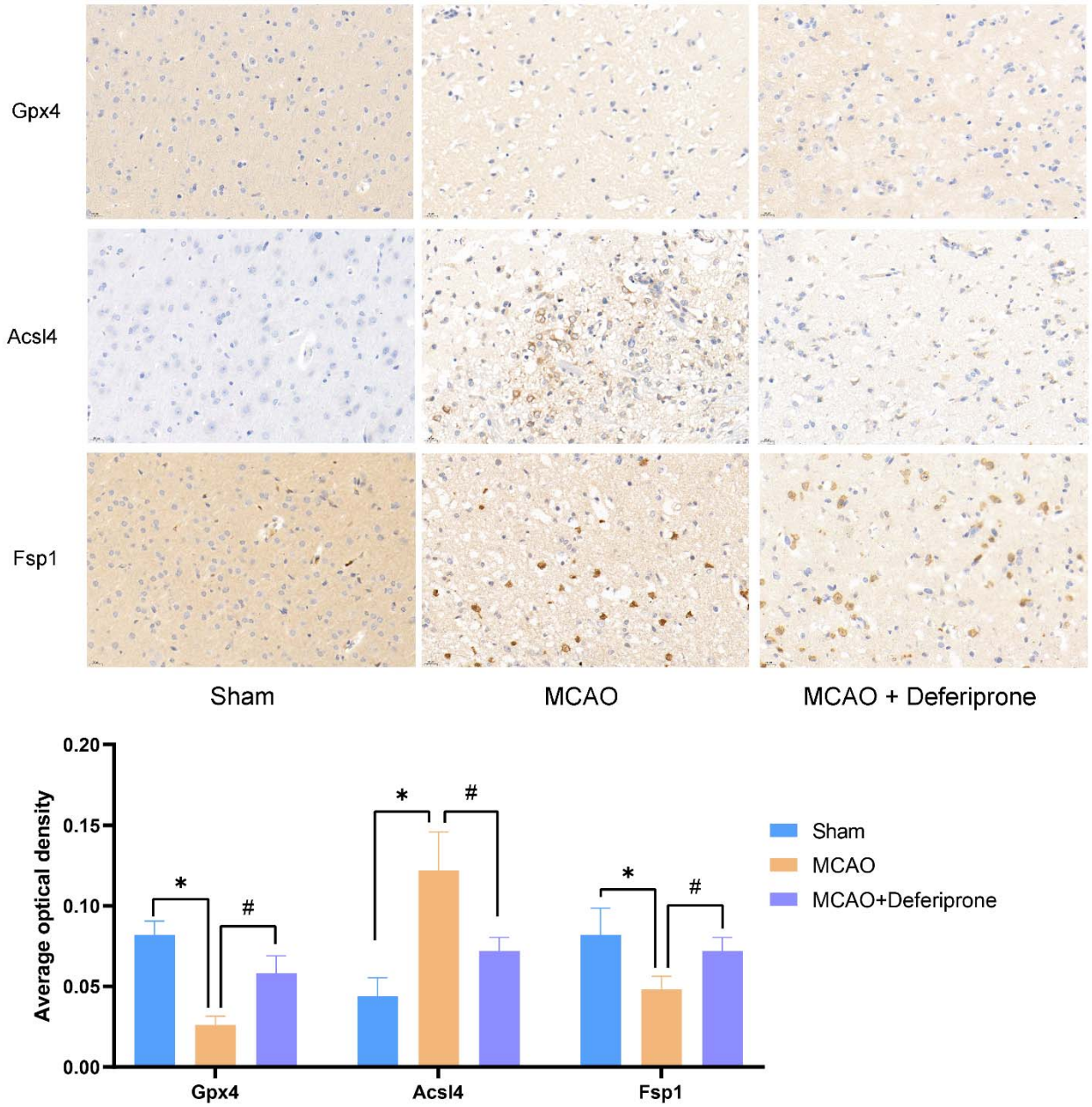

Figure 10 Expression levels of Acsl4 and Gpx4 proteins. Immunohistochemistry, 40× magnification. *compared with the sham group, P<0.05; #compared with MCAO group, #P<0.05. n=3 per group

2.6.4 Expression levels of *Bax, IkBa, Casp3, p-Smad, p65(Rela), Tfrc, Trf and Sirt1* proteins detected by immunofluorescence staining

Compared to the Sham operation group, the MCAO group exhibited significantly elevated levels of Bax, IκBα, p-Smad, p65 (Rela), Tfrc, and Trf proteins (P<0.05). These results are consistent with the differential gene expression findings from the single-cell sequencing analysis. Meanwhile, Compared to the Sham operation group, the Casp3 protein level was increased, while the Sirt1 protein level was decreased in the MCAO group (P<0.05) (Figure 11).

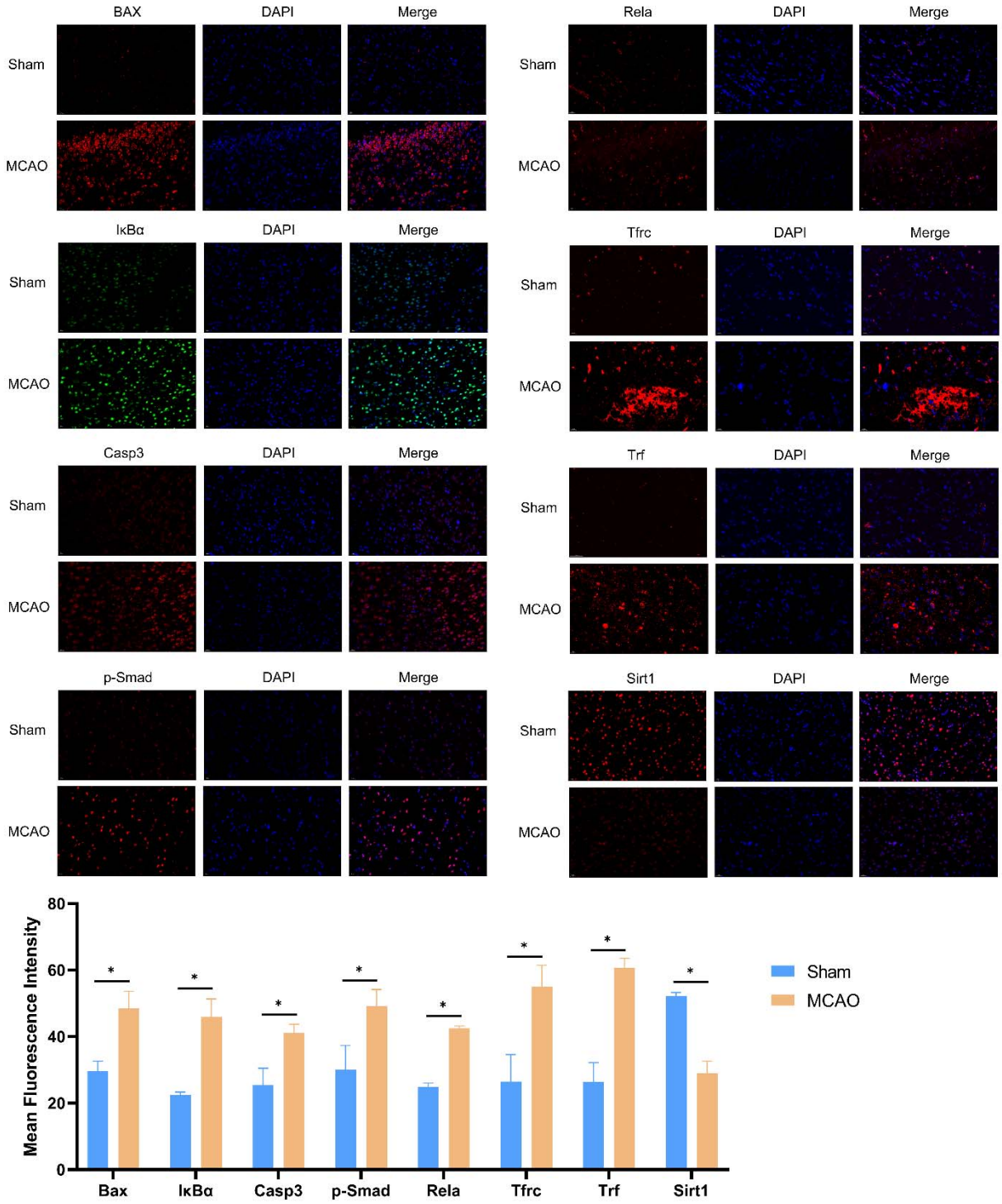

Figure 11 Expression levels of Bax, IkBa, p-Smad, p65, Tfrc, and Trf proteins. Immunofluorescence, 40× magnification. Compared with the sham group, *P<0.05. n=3 per group.

2.7 Identifying Epigenetic Metabolic Regulation in NSCs Using Novel Functional Mapping

In Section 2.6, we validated the algorithm's effectiveness through animal experiments and conducted biological discoveries in regenerative medicine. Kremer et al. demonstrated that astrocytes

and quiescent NSCs exhibit highly similar transcriptomic profiles despite their distinct functional roles[2]. However, research suggests that under specific conditions—such as injury or overexpression of key factors—astrocytes can reactivate their latent neurogenic potential. This finding offers new possibilities for regenerative medicine: by regulating cellular memory and energy metabolism, it may be possible to induce astrocytes' neurogenic capacity, enabling regenerated organs to acquire neuronal activity. Effective organ repair and regeneration require precise control of cellular energy states, yet the molecular mechanisms encoding cellular energy supply remain unknown. Glial cells are known to exist in distinct active and quiescent states, and astrocyte activity may be reflected in metabolic changes, with cellular memory governing the initiation of energy metabolism. Cellular memory involves multiple factors, including protein, RNA, and DNA modifications, but the mechanisms sustaining cell activity remain unclear. Although astrocytes and quiescent NSCs differ functionally, their transcriptomic similarities have traditionally required expensive single-cell identification techniques to capture activation events and distinguish their states. This challenge extends to broader applications. For instance, improving energy metabolism—such as enhancing mitochondrial function or supplying metabolic substrates—can significantly reduce myofibroblast activation. This approach holds potential for reversing heart failure by promoting the transition of myofibroblasts into fibroblasts, ultimately restoring cardiac function[15].

In a recent study, Kremer et al. utilized DNA methylation and RNA transcriptional data to uncover the stem-like properties of astrocytes under ischemic conditions[2]. They stratified the NSCs and their progeny as illustrated in Figure 12. Among them, qNSCs can differentiate into activated neural stem cells (aNSCs). qNSCs mainly include ventricular–subventricular zone (vSVZ) astrocytes [Here referring to dormant NSCs (qNSC1), deeply quiescent neural stem cells in the vSVZ) and qNSC2 (Primed quiescent neural stem cells). aNSCs can differentiate into transient amplifying progenitors (TAPs) and further differentiate into neuroblasts.

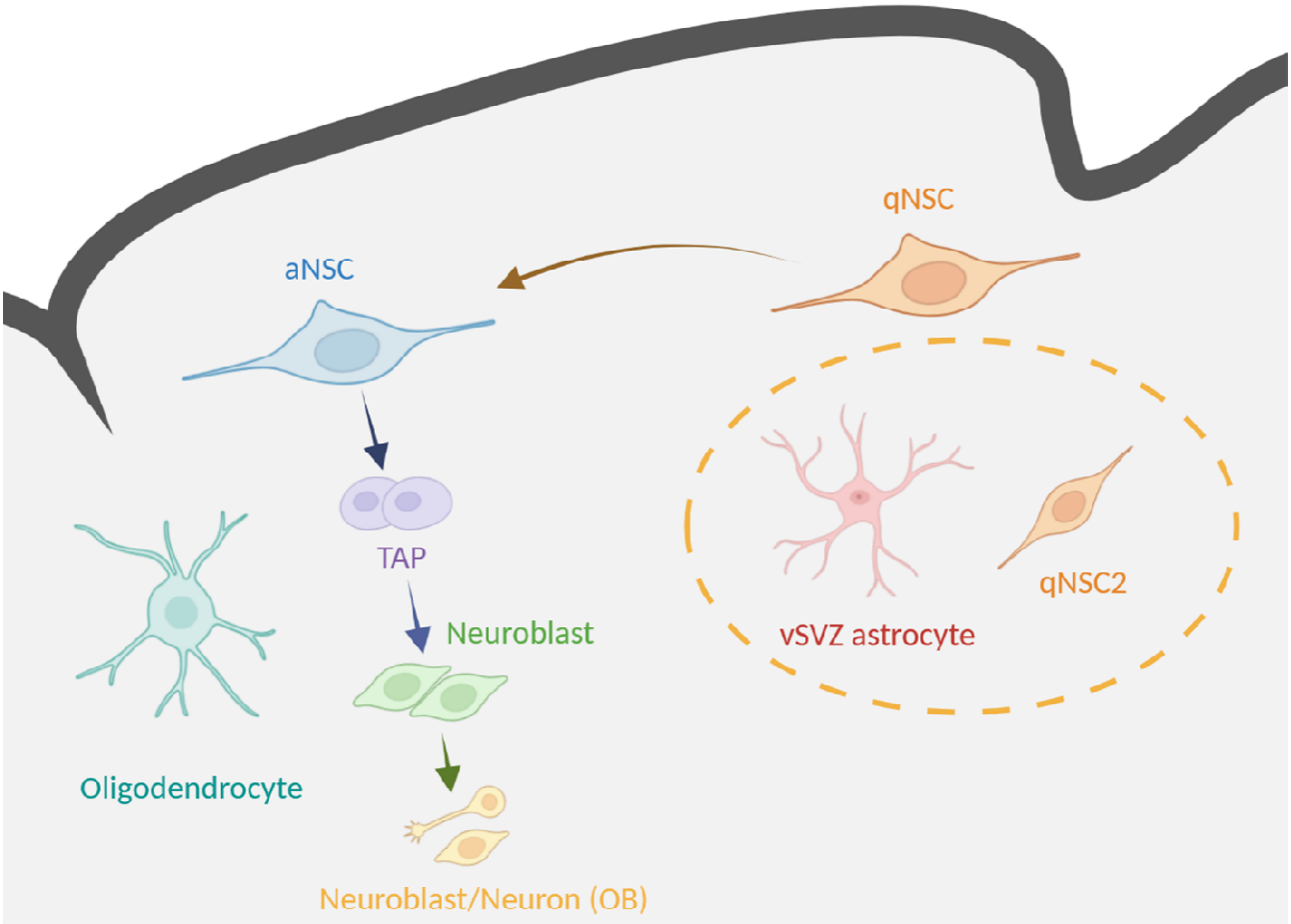

Figure 12 Schematic diagram of the vSVZ adult NSC lineage. Referenced from Kremer et al. 2024[2]. Created by Biorender. qNSC: quiescent neural stem cell; qNSC2: Primed quiescent neural stem cells; TAP: transient amplifying progenitors; OB: olfactory bulb

Transcriptome hierarchical clustering groups qNSC1 and qNSC2 cells with astrocytes from other tissues (Figure 13b). In stark contrast, methylome clustering only groups qNSC1 cells with astrocytes from other regions, while qNSC2 cells cluster with other cells in the NSC lineage. This indicates that in quiescent NSC cells (qNSC1), NSC cells share more similar transcriptional states with astrocytes, and since transcriptional states are controlled by current epigenetic memory, qNSC1 possesses epigenetic memory more similar to astrocytes. When NSC cells transition to qNSC2 cells, the methylome of qNSC2 cells is reprogrammed to the NSC methylome, resulting in an alteration of epigenetic memory.

The success of this study is based on two facts: (1) The need to know in advance the two quiescent states of NSC cells, namely qNSC1 and qNSC2. (2) The existence of two states available for modeling. It is necessary to infer which are Astrocytes, and to know in advance which are Neuroblasts. The original authors identified the real label of cells.

We used TOGGLE to analyze the GSE209656 dataset tested by the authors. Cells appearing in the same group are those with similar gene expression. For example, in the first round of classification, we identified astrocytes (striatum), neuroblasts, oligodendrocytes, reactive astrocytes, TAPs and vSVZ astrocytes. We also identified astrocytes (striatum), a mixed state of astrocytes (cortex) and vSVZ astrocytes, which we named For_R2-1; and a mixed state of astrocytes (striatum) and vSVZ astrocytes, which we named For_R2-2; we also found some cells mixed together, named Mixed cells. These cells have similar gene expression profiles. The fact that these cells are classified into the same group suggests that they may have similar epigenetic memory, leading to similar behavioral patterns. This point has been verified in the original text of the author. Our calculation results (Figure 13a) are basically consistent with the labels provided by Kremer et al. (Figure 13b).

Then, we used TOGGLE to further analyze For_R2-1 and For_R2-2. For For_R2-1, as shown in Figure 13d, we divided the cells into 3 groups based on the calculation results, which is basically consistent with the labels provided by Kremer et al. (Figure 13e). For For_R2-2, as shown in Figure 13f, we divided the cells into 2 groups based on the calculation results, which is basically consistent with the true labels (Figure 13g). In summary, the classification results are summarized in Figure 13c. This method can also be used to classify and identify cells in Mixed cells.

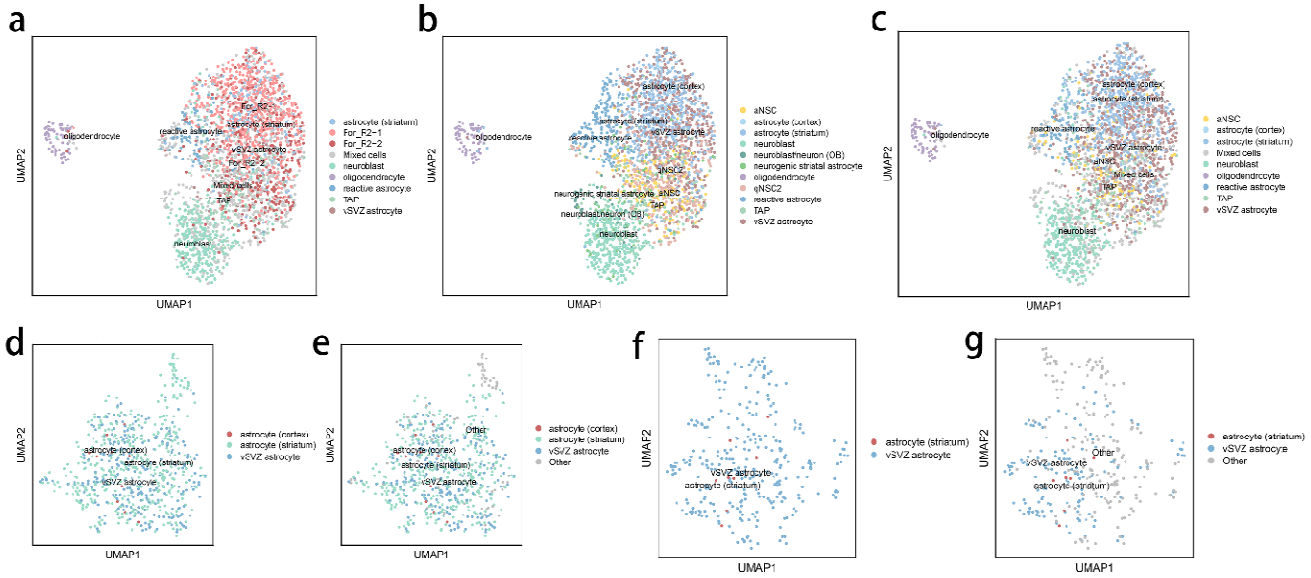

Figure 13 Results of using TOGGLE to identify the glial cell GSE209656 dataset. a: Cell populations after the first round of characterization; b: Labels provided by Kremer et al.[2]; c: Overall identification of cell populations; d: Cell populations identified for For_R2-1; e: Labels of the For_R2-1 cell population provided by Kremer et al.[2]; f: Cell populations identified for For_R2-2; g: Labels of the For_R2-2 cell population provided by Kremer et al.[2]

2.7.1 Establish Graph Diffusion Functional Map

We used UMAP to visualize the functional grouping of RNA. While UMAP effectively captures overall RNA expression trends, the dimensionality reduction process introduces information loss, leading to mixed functional groupings. This mixing effect makes it difficult to accurately distinguish RNA groups based on functional states, thereby limiting the depth of target identification and functional analysis.

To address this issue, we developed an innovative Graph Diffusion Functional Map. As detailed in Section 4.2, replacing the cell-based graph diffusion matrix M_{ij} with an RNA-based matrix and applying the full methodology outlined in the Methods section generates the functional map. This approach preserves the original RNA correlation distribution while significantly reducing functional mixing, allowing for clearer visualization of RNA functional groupings. It also enhances the detection of subtle functional differences in high-dimensional data without the dimensionality loss inherent in traditional reduction methods. Our results revealed exciting discoveries: the Graph Diffusion Functional Map not only validated known enriched clusters but also identified previously uncharacterized RNA groups with significantly high expression levels. These potential targets, often obscured or overlooked in conventional UMAP analyses, were clearly distinguished in our functional map. This breakthrough offers a novel perspective for single-cell transcriptomics analysis and expands possibilities for studying disease mechanisms and identifying therapeutic targets.

We further applied TOGGLE to classify functional RNA in ferroptosis-associated cell populations (Group R3-4 and R3-5) from Section 2.5.3, as well as to the top 1,000 highly variable genes in aNSCs and qNSCs from the GSE209656 dataset. The results were visualized using our method (Figure 14).

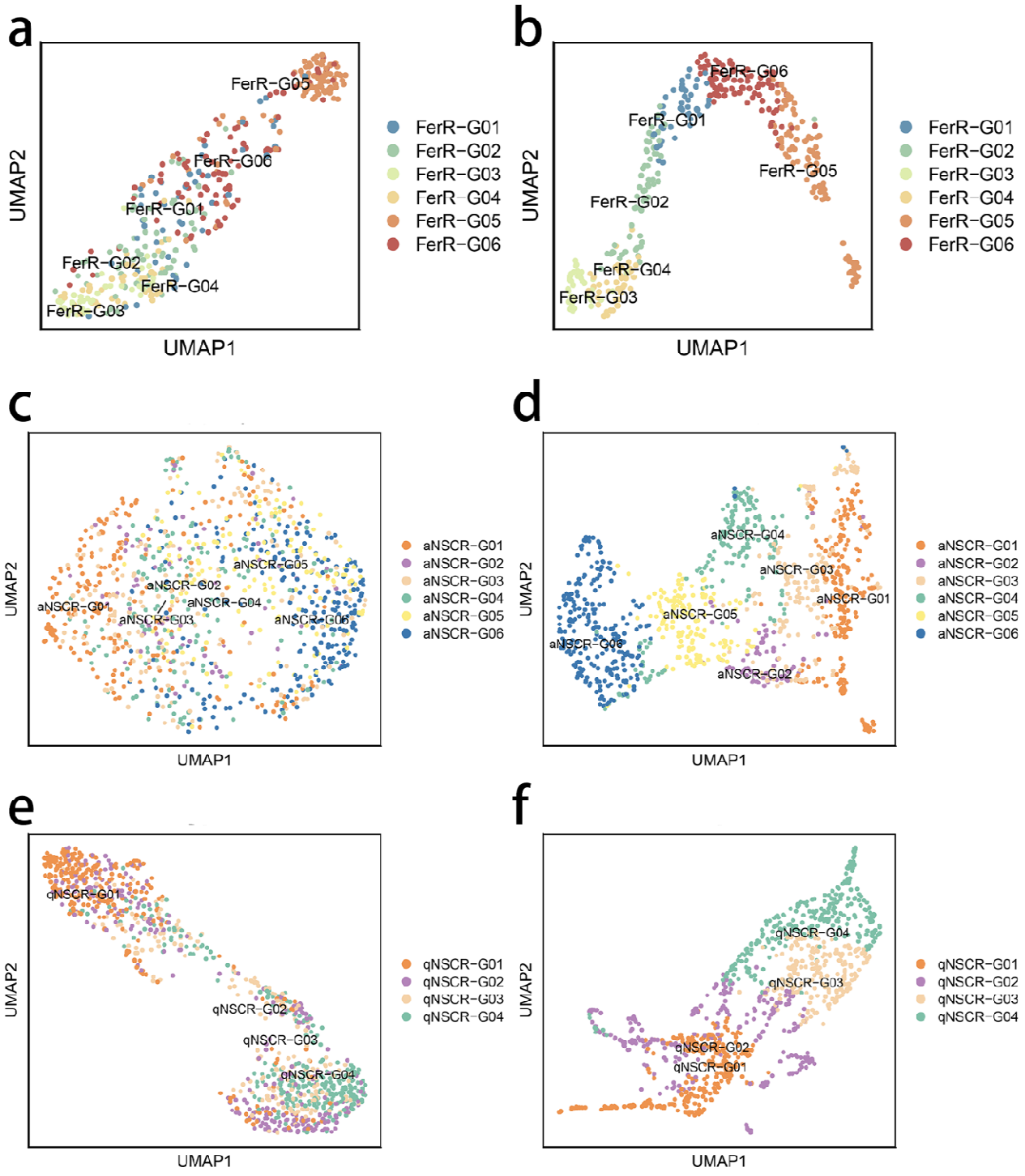

Figure 14 RNA functional profiles from different data. a: RNA map of ferroptosis-related genes in Group R3-4 and R3-5 cells made by using single-cell expression matrix; b: RNA map made using the new method for the data in a; c: RNA map of 1000 highly variable genes in aNSC cells made by using single-cell expression matrix; d: RNA map made using the new method for the data in c; e: RNA map

of 1000 highly variable genes in qNSC cells made by using single-cell expression matrix; f: RNA map made using the new method for the data in e

2.7.2 Functional discovery of RNA groups

We performed enrichment analysis on the RNA groupings presented in Figure 14a and 14b. The enrichment analysis results of FerR-G01 include cellular response to starvation (GO:0009267), cellular response to hypoxia (GO:0071456), positive regulation of blood vessel endothelial cell migration (GO:0043536), and positive regulation of cell migration (GO:0030335). These suggest that this RNA cluster may play a role in promoting angiogenesis in ischemic brain regions to counteract hypoxia caused by cerebral ischemia. Within this RNA group, we identified ferroptosis-related genes Nfs1, Osbpl9, and Pgd, which were not enriched in any of the related biological processes, implying their functions may be previously unrecognized. We hypothesize that these genes are functionally linked to the processes associated with FerR-G01.

FerR-G02 is associated with negative regulation of apoptotic processes (GO:0043066) and cellular response to oxidative stress (GO:0034599), indicating that this RNA cluster may be involved in both ferroptosis-related oxidative stress and apoptosis. Notably, several ferroptosis-related genes, Pvt1, Abhd12, Cbr1, Far1, and Sat1, were not enriched in the identified processes, suggesting their potential involvement in oxidative stress regulation.

FerR-G03 is associated with autophagy (GO:0006914), autophagosome maturation, and negative regulation of transcription by RNA polymerase II (GO:0000122), indicating a possible role in both ferroptosis and autophagy regulation. The genes Gstm1, Rpl8, Ubc, and Vdac2 within this group may also be functionally relevant to these processes.

FerR-G04 is linked to cell redox homeostasis (GO:0045454), regulation of the cell cycle (GO:0051726), and the endoplasmic reticulum unfolded protein response (GO:0030968), suggesting its involvement in ferroptosis-related oxidative stress. The genes Agpat3, Abcc5, Cd82, Emc2, Cs, and Psat1 within this cluster may also contribute to these functions.

FerR-G05 is associated with superoxide metabolic processes (GO:0006801), inflammatory responses (GO:0006954), and cellular response to hydrogen peroxide (GO:0070301), implying a connection to oxidative stress in ferroptosis. The genes Amn, Cdca3, Cltrn, Hcar1, Plin4, and Galnt14 may also play roles in these pathways.

FerR-G06 is associated with cellular response to reactive oxygen species (GO:0034614) and SMAD protein signal transduction (GO:0060395), suggesting a link between ferroptosis-related oxidative stress and the SMAD signaling pathway. Several genes within this cluster, including Cisd3,

Hddc3, Lyrm1, Rbms1, Ahcy, Adam23, Aqp5, Ccdc6, Eno3, Gria3, Gpt2, Gstz1, Ints2, Klhdc3, Lpcat3, Mib2, Phkg2, Parp4, and Pdss2, may also be functionally relevant (Figure 15).

The genes not enriched in the identified signaling pathways may still be functionally relevant to their respective RNA clusters, potentially revealing new mechanisms in ferroptosis. For example, in FerR-G01, Nfs1 plays a crucial role in the biosynthesis of iron-sulfur (Fe-S) clusters, which are essential for the function of mitochondrial respiratory chain complexes[16]. Ischemia-induced hypoxia disrupts mitochondrial function, reducing ATP production and causing cellular damage. Nfs1 may regulate ferroptosis by maintaining Fe-S cluster homeostasis, thereby modulating iron metabolism and antioxidant capacity. Osbpl9 is a key regulator of cholesterol and phospholipid exchange, playing an important role in intracellular lipid transport and signaling[17]. It may promote endothelial cell migration and neovascularization in ischemic brain regions by modulating the lipid microenvironment and associated signaling pathways. Pgd participates in the pentose phosphate pathway (PPP), primarily generating NADPH to maintain cellular antioxidant capacity[18]. Under hypoxic conditions, elevated reactive oxygen species (ROS) levels contribute to oxidative stress. Pgd may counteract ferroptosis by enhancing antioxidant defenses, protecting cells from oxidative damage.

In FerR-G02, Pvt1 is a long non-coding RNA (lncRNA) known to regulate various physiological processes, including cell proliferation, apoptosis, and oxidative stress response[19]. In the context of ferroptosis, Pvt1 may influence iron accumulation and ROS generation by modulating stress-response genes. Abhd12, an enzyme involved in lipid metabolism, plays a key role in membrane lipid processing [20]. It may contribute to ferroptosis resistance by regulating fatty acid metabolism, stabilizing cellular membranes, and modulating intracellular lipid homeostasis. Cbr1 is an oxidoreductase enzyme that catalyzes the reduction of aldehyde compounds, protecting cells from oxidative damage [21]. It may mitigate ferroptosis by reducing ROS levels and limiting oxidative injury. Far1 is involved in peroxisome biogenesis[22], suggesting a potential role in oxidative stress regulation and ferroptosis modulation. Sat1, a key gene in the polyamine metabolism pathway, facilitates polyamine acetylation, which regulates intracellular polyamine levels[23]. Given that polyamines exhibit antioxidant properties, Sat1 may counteract oxidative stress, thereby influencing ferroptosis susceptibility.

Additionally, we performed enrichment analysis on the RNA clusters derived from the top 1,000 highly variable genes in aNSCs, as shown in Figures 14c and 14d. Our findings indicate that aNSCR-G01 is primarily associated with cell differentiation (GO:0030154), while aNSCR-G02 is linked to the regulation of transcription by RNA polymerase II (GO:0006357). aNSCR-G03 is enriched in chromosome segregation (GO:0007059) and proteolysis (GO:0006508), whereas aNSCR-G04 is involved in DNA repair (GO:0006281) and cell division (GO:0051301). aNSCR-G05 is related to cell

migration (GO:0016477), regulation of gene expression (GO:0010468), and positive regulation of cell growth (GO:0030307), while aNSCR-G06 is enriched in mRNA processing (GO:0006397), RNA splicing (GO:0008380), and chromatin remodeling (GO:0006338). These functional associations suggest that the identified RNA clusters are predominantly involved in cell proliferation, DNA replication, and RNA transcription, aligning well with the in vivo characteristics of aNSCs.

We also performed enrichment analysis on the RNA clusters derived from the top 1,000 highly variable genes in qNSCs, as shown in Figures 14e and 14f. Our findings indicate that qNSCR-G01 is primarily associated with centriole replication (GO:0007099) and regulation of the mitotic cell cycle (GO:0007346), while qNSCR-G02 is linked to cell adhesion (GO:0007155) and the apoptotic process (GO:0006915). qNSCR-G03 is enriched in regulation of transcription by RNA polymerase II (GO:0006357) and endoplasmic reticulum to Golgi vesicle-mediated transport (GO:0006888), whereas qNSCR-G04 is involved in proton motive force-driven mitochondrial ATP synthesis (GO:0042776) and proton transmembrane transport (GO:1902600). These functional associations suggest that the identified RNA clusters are involved in regulating the cell cycle, cell adhesion, intracellular protein synthesis, and cellular energy metabolism—characteristics typical of non-proliferative cells, aligning well with the in vivo behavior of qNSCs.

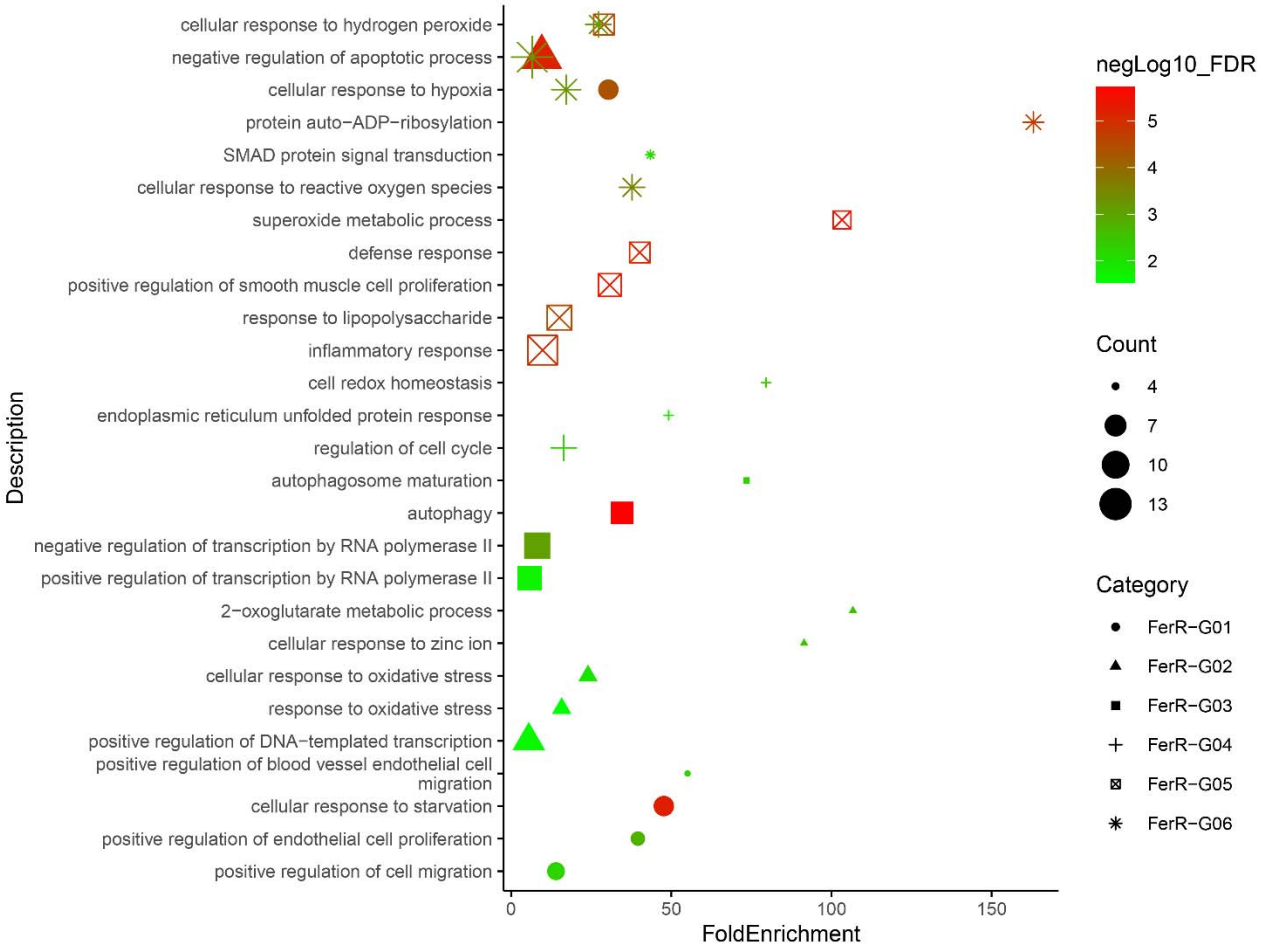

Figure 15 Bubble chart of biological processes of ferroptosis-related genes in Group R3-4 and R3-5 cells

2.7.3 Discovery of epigenetic control of energy metabolism

To investigate the epigenetic regulation of energy metabolism, we conducted a multimodal analysis of NSC scRNA-seq and single-cell epigenomic data from Kremer et al.[2] First, we extracted aNSCs and qNSC2 based on the labels provided by Kremer et al.[2] and analyzed the data of GSE209656 (scRNA-seq data) using TOGGLE (Figure 16a and 16b). Next, we classified the cells into two groups based on the median expression level of mitochondrial RNA: a high mitochondrial RNA expression group (High-MT) and a low mitochondrial RNA expression group (Low-MT) (Figure 16d and 16e). Our analysis revealed that NeuG04 primarily consists of metabolically active cells, predominantly aNSCs, while NeuG01 is composed mainly of metabolically inactive cells, predominantly qNSCs (Figure 16c and 16f). The remaining cell groups fall between these two cell groups and may represent transitional states.

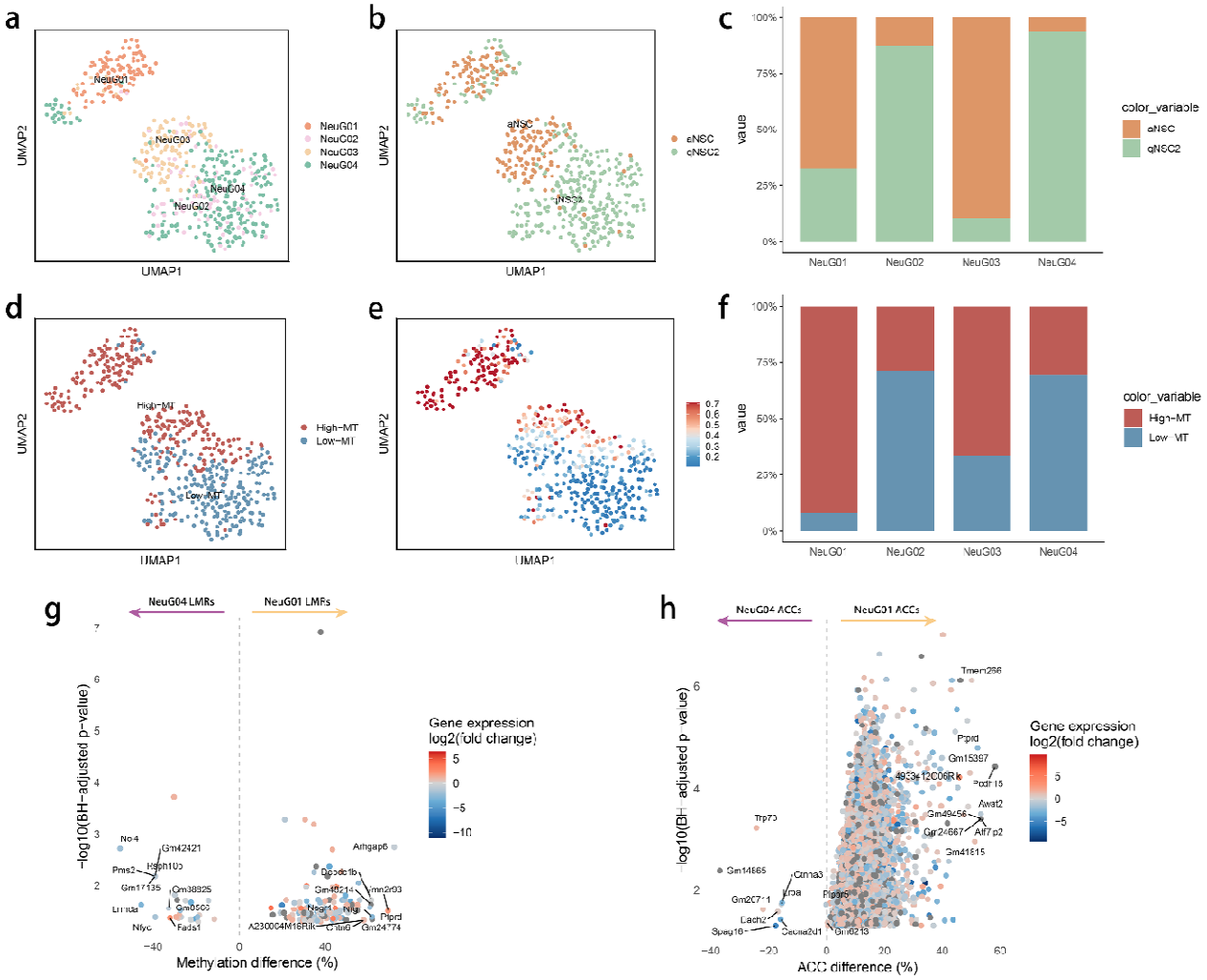

Figure 16 Results of epigenetic analysis of aNSCs and qNSCs. a: Cell populations after TOGGLE analysis; b: Labels provided by Kremer et al.; c: The proportion of aNSC and qNSC2 in each cell group; d: Cells are divided into High-MT and Low-MT according to the expression of mitochondrial genes; e: Mitochondrial gene expression in cells; f: The proportion of high-MT cells and low-MT cells in each cell group; g: Volcano plot of LMRs, tested for differential methylation between NeuG01 and NeuG04 (two-sided Wilcoxon test), LMRs are colored according to the differential expression of the nearest gene. h: Volcano plot of high ACC resgions, tested for differential ACC between NeuG01 and NeuG04 (two-sided Wilcoxon test), LMRs are colored according to the differential expression of the nearest gene

Then, we analyzed the single-cell epigenomes (GSE211786) dataset from Kremer et al.[2] based on the cell grouping results to analyze the epigenetic differences between the high-MT-dominated cell population (NeuG01) and the low-MT-dominated cell population (NeuG04). First, we extracted the demethylation data for cells classified as NeuG01 and NeuG04 and analyzed the variably methylated

regions (VMRs) using Kremer et al.'s approach. We then calculated average DNA methylation (meth) and chromatin accessibility (ACC).

To identify differentially methylated regions between the two groups, we performed a two-sided Wilcoxon rank sum test on the VMRs, considering only regions with genomic coverage in at least six cells per group. VMRs with a Benjamini–Hochberg-adjusted P-value < 0.05 were labeled as lowly methylated regions (LMRs) or high ACC regions. To annotate the genes, all LMRs overlapping a gene body were assigned to that gene.

The results revealed that compared to NeuG04, NeuG01 exhibited higher LMR and high ACC regions (Figure 16g and 16h). The LMRs in NeuG01 were enriched in genes associated with stem cell differentiation (GO:0048863) and positive regulation of transcription by RNA polymerase II (GO:0045944), while high ACC regions in NeuG01 were linked to genes involved in cell adhesion (GO:0007155), chromatin remodeling (GO:0006338), axon guidance (GO:0007411), positive regulation of transcription by RNA polymerase II (GO:0045944), actin cytoskeleton organization (GO:0030036), nervous system development (GO:0007399), cell migration (GO:0016477), positive regulation of DNA-templated transcription (GO:0045893), and cell differentiation (GO:0030154). These findings suggest that the epigenetic modifications in NeuG01 cells are primarily associated with cell proliferation, differentiation, and migration—processes that demand high energy consumption.

Conversely, the LMRs in NeuG04 were linked to genes involved in intracellular signal transduction (GO:0035556), nervous system development (GO:0007399), regulation of synapse assembly (GO:0051963), cell adhesion (GO:0007155), synaptic membrane adhesion (GO:0099560), signal transduction (GO:0007165), long-term memory (GO:0007616), negative regulation of the cell cycle (GO:0045786), negative regulation of cell migration (GO:0030336), and negative regulation of DNA-templated transcription (GO:0045892). These functions are negatively correlated with cell proliferation and division, indicating that NeuG04 cells are transitioning into a quiescent state. A detailed list of LMRs, high ACC regions, and their corresponding genes for NeuG01 and NeuG04 is provided in Table S4.

Since NeuG01 is primarily composed of aNSC cells, while NeuG04 consists mainly of qNSC2 cells, we can conclude the following: (1) During the transition from qNSC2 to aNSC, key genomic regions associated with cell proliferation and energy metabolism undergo significant demethylation (Figure 16g), accompanied by a notable increase in chromatin accessibility (Figure 16h). (2) In cell populations dominated by aNSCs, the proportion of high mitochondrial gene expression (High-MT) cells is significantly higher than in qNSC2 (Figures 16c, 16d, and 16e), further indicating that aNSCs exhibit higher metabolic activity.

In summary, we propose that during the qNSC2-to-aNSC transition, localized DNA demethylation and chromatin opening activate the expression of genes involved in cell proliferation and energy metabolism, leading to a gradual increase in metabolic activity. This epigenetic modification reflects differences in “cellular memory” between these two cell types and may be a key mechanism underlying their distinct mitochondrial energy metabolism states.

3 Discussion

This study introduces TOGGLE, an unsupervised learning method based on scRNA-seq data for distinguishing cellular functions. Our algorithm integrates concepts from deep learning and deep encoding to enhance analytical capabilities. In Section 4.2, "Graph Diffusion," we propose a deep learning-based approach for adapting unsupervised frameworks to supervised tasks[24]. This method addresses challenges associated with asymmetric graph structures in cellular relationships and exhibits strong adaptability to heterogeneous data, where nodes and edges have varying types or weights, effectively replacing traditional supervised deep learning methods. This step is primarily used to compute functional similarities between cells. Self-supervised learning (SSL) is a specialized form of unsupervised learning. In Section 4.3, "Calculate Correlation Metrics," we embed graph diffusion within an environment-responsive self-supervised design, an innovative approach[25]. In this method, self-supervised learning does not rely on manually labeled data but instead generates correlation labels directly from the data. These labels are subsequently used as additional environmental conditions to iteratively refine new labels. This mechanism is inspired by reinforcement learning, where features and rules are autonomously discovered through environmental feedback. This step is primarily used to cluster cells into distinct functional groups.

Bidirectional Encoder Representations from Transformers (BERT) is a pretrained model based on a bidirectional Transformer encoder. It learns deep semantic relationships in context through Masked Language Modeling (MLM) and is widely used in text generation tasks[26]. It also plays a crucial role in large language models (LLMs) such as ChatGPT, Claude, and Gemini for encoding vector generation. In Section 4.5, "Deep Genetic Network," we introduce a novel unsupervised deep learning network that incorporates a distance optimization framework to reduce errors in self-supervised learning. To enhance the masking mechanism, we employ subspace reconstruction, redesigning the traditional masking approach[27]. This redesign enables the masked language framework to process numerical data. Similar to BERT's masking mechanism, in this network, each cluster's true labels are treated as masked language information, and each iteration of the genetic network generates a new subspace. According to Takens' Theorem (provided in the appendix), multiple encoded outputs generated through iterative subspaces are

stacked to form a feature vector. With each additional subspace, one mask is removed, gradually expanding the original encoding's feature representation. The fixed encoding portion preserves real spatial information, while the progressive generation of new spatial information gradually removes masks over iterations. Decoding the final feature vector yields predicted cluster classifications. Ultimately, clusters with identical predicted classifications are merged to obtain the final true cell classification results.

With the functional gene expression definition tools described above, we can conduct more detailed genetic lineage comparison analyses, which not only deepens our understanding of the human genome but also opens new avenues for applying functional gene regulation in immunotherapy and organ regeneration. For instance, in cardiac regeneration, ensuring consistency in gene expression across in vivo and in vitro stages is crucial for obtaining usable regenerative organs. This requires carefully designed synchronized experimental environments, including segmented cultures to select cells with consistent functional patterns. Furthermore, using intercellular communication studies to assess the immune response in regenerating tissues is necessary. This evaluation must be based on pre-defined functional clusters to accurately judge the immune effects produced by similar cell types. These applications of human genetic lineages span several key fields, from cancer treatment to changes in cell function during heart failure. Additionally, viral infections can lead to systemic changes in cellular RNA expression, necessitating functional comparison analysis between cells before and after infection, which essentially studies the functional transitions of the same cell type. It is important to emphasize that this task is easily confused with pseudotime calculations. The goal of pseudotime analysis is to recover temporal information from fate mapping, without distinguishing between different stage boundaries. In contrast, fate tracking not only recovers temporal information but also clusters cells with similar states, delineating boundaries for different phases within the same programmed fate trajectory. This approach further reveals intercellular communication relationships and functional clustering.

With the functional gene expression delineation tools established above, we can now conduct more refined gene lineage comparison analyses. This not only deepens our understanding of the human genome but also paves the way for applying functional gene regulation in immunotherapy and organ regeneration. For instance, in cardiac regeneration, ensuring consistency between in vivo and in vitro gene expression stages is crucial for obtaining viable regenerative organs. Achieving this requires carefully synchronized experimental environments and segmented culturing to filter cells with matching functional patterns. Additionally, intercellular communication studies are essential for evaluating the immune response of regenerated tissues. This evaluation must be based on predefined functional clusters to accurately determine the distinct immune impacts produced by cells of the same type. The application

of human gene lineage analysis extends across several critical fields, from cancer therapy to functional state changes in heart failure. Moreover, viral infections can induce systematic alterations in cellular RNA expression, necessitating a comparative functional analysis of cells before and after infection. Fundamentally, this remains a study of functional transitions within the same cell type, highlighting the broader relevance of these analytical frameworks. Compared with mainstream algorithms, our method demonstrates the following advancements and research differences, establishing TOGGLE as a novel computational framework (Table 1 and Table 2):

Table 1: Advanced algorithms

Previous	Our Advantages	Reference
Cell Rank	Used for grouping cell subtypes. TOGGLE not only enables subtype grouping but also distinguishes functional stages and fate stages.	Marius Lange et al.[28] *Nature Methods*
DeepCell	Pre-training is required. TOGGLE does not require pre-training. In Fig. 1, we introduce Problems 1 and 4, which are typically not applicable to pre-training scenarios.	Dylan Bannon et al.[29] *Nature Methods*
scGen	Cell perturbation and infection responses across different cell types. TOGGLE can be used to analyze cell perturbation and functional differences within the same cell type under different states.	Mohammad Lotfollahi et al.[30] *Nature Methods*
CellTypist	Previous studies focused on cell subtype classification and cell type classification. Building on this foundation, TOGGLE can further distinguish different functional states within the same cell subtype, as well as differentiate the programmed fate stages within the same cell subtype.	C. Domínguez Conde et al.[31] *Science*
SCENIC		Sara Aibar et al.[32] *Nature Methods*
pySCENIC		Bram Van de Sande et al.[33] *Nature Protocols*
Waddington-OT	Cell subtype classification. TOGGLE can perform developmental and functional differentiation, as well as programmatic fate stage distinction within the same subtype.	Geoffrey Schiebinger et al.[34] *Cell*
Markov chain Models (MARGARET)	Used to identify how many developmental outcomes a cell has. TOGGLE not only distinguishes multiple developmental outcomes but also performs subgroup classification, developmental prediction, program fate stage categorization, and differentiation of cells with different functions within the same subtype.	Brian K. Chu[35] *BMC Systems Biology*
Cospar	Used for binary classification of progenitor cells. TOGGLE can be applied to multi-class classification of both progenitor and mature cells.	Wang et al.[7] *Nature Biotechnology*

Table 2: Research differences

Algorithms	Subtype	Development		Communication	Death	Functional RNA	Functional regulation distinction		
	Cell subtype differentiation	Developmental prediction	Distinguishing between progenitor and daughter cells	Functional classification of intercellular communication	Label-free identification of cells with promiscuous shared pathways	RNA functional cluster classification	Epigenetic memory distinction, taking methylation as an example	RNA transcription activity changes in epigenetic memory	Differentiate between the developmental stages of the same cell
Cell Rank	√	×	√	×	×	×	Part*1	Part*1	×
Deep Cell	√	×	√	×	Part*2	×	×	×	×
scGen	√	×	×	Part*1	×	×	×	×	×
CellTypist	√	×	√	Part*1	×	×	×	×	×
SCENIC	√	×	×	×	×	Part*1	×	×	×
pySCENIT	×	×	×	Part*4	×	Part*4	×	×	×
WOT	√	√	√	×	×	×	×	Part*3	×
Cospar	×	√	×	×	×	×	×	×	×
MARGARET	√	×	×	×	×	×	×	×	×
TOGGLE	√	√	√	√	√	√	√	√	√
Pseudo-time type algorithm	×	×	×	×	×	×	×	×	×

Part*1 is limited to distinguishing between different cell types; Part*2 can only identify cells that have undergone terminal cell death; Part*3 requires a fully comprehensive reference dataset; Part*4 is restricted to analyzing regulatory networks.

For the NSC scRNA-seq and single-cell epigenomic sequencing datasets, we classified cells into High-MT and Low-MT groups based on mitochondrial RNA expression levels in scRNA-seq data. Using TOGGLE, we then analyzed aNSC and qNSC2 scRNA-seq data, clustering the cells into four distinct groups. Among them, we extracted NeuG01 and NeuG04 for further epigenomic analysis. Our findings revealed significant epigenetic differences between NeuG01, predominantly composed of High-MT and aNSC cells, and NeuG04, primarily consisting of Low-MT and qNSC cells. Based on these results, we hypothesize that during the qNSC2-to-aNSC transition, localized DNA demethylation and chromatin opening activate the expression of genes involved in cell proliferation and energy metabolism, leading to a gradual increase in metabolic activity. This epigenetic modification reflects differences in “cellular memory” between these two cell types and may serve as a key mechanism underlying their distinct mitochondrial energy metabolism states. Overall, this demonstrates that our tool, TOGGLE, enables multimodal analysis of scRNA-seq and single-cell methylation data, allowing for the identification of subtle differences within the same cell type.

By utilizing TOGGLE to analyze neuronal data after ischemic stroke (GSE232429), we successfully mapped the trajectory from healthy neurons to those undergoing programmed cell death. Differential gene expression analysis revealed downregulation of *Gpx4* and upregulation of *Trf*, *Tfrc*, *Smad1*, *Smad5*, *Acsl4*, *Rela*, *Bax*, and *Nfkbia* in neurons experiencing programmed cell death compared to healthy counterparts. This trend was validated in animal models of ischemic stroke, further confirming TOGGLE's utility in constructing fate trajectories. In addition, several apoptosis- and ferroptosis-related genes showed expressions inconsistent with previous studies. For example, Aifm2 and *Sirt1* are generally downregulated during ferroptosis[36,37] but were upregulated in dying neurons in the GSE232429 dataset, whereas *Casp3* is upregulated during apoptosis[38], but was downregulated in dying neurons in GSE232429. Our experiments also found that Aifm2 and Sirt1 were downregulated, while Casp3 was upregulated in the MCAO group. We infer that the reason for this phenomenon may be that cells in the early stages of programmed cell death are trying to "self-rescue". Future tests are needed to capture dying cells to confirm this hypothesis. In the future, TOGGLE may provide insights into genetic differences between cells at various stages of programmed cell death, such as apoptosis and necroptosis. This could lead to the identification of novel biomarkers and pathways, potentially guiding the development of targeted intervention drugs aimed at modulating these processes.

Among the forms of programmed cell death, ferroptosis—a process driven by iron-dependent lipid peroxidation—has emerged as particularly relevant[39]. Ferroptosis is now recognized as a key mechanism underlying ischemic stroke, with inhibition of this pathway offering a promising therapeutic approach. However, the incomplete understanding of ferroptosis mechanisms has limited advances in treatment

exploration[5]. Our study not only mapped the trajectory of neuronal programmed cell death following ischemic stroke but also identified neuron populations primarily undergoing ferroptosis. We also found some positively expressed ferroptosis-driving genes among the differentially expressed genes, such as *Ctsb, Mtdh, Ndrg1, Smad7, Egr1, Agpat3, Mapk3, Mdm2, Cd82, Lgmn, Elavl1, Nras, Mdm4, Lpin1, Lig3, Idh1, Por, and Acsl1* (Table S2). Targeted intervention of these genes in the future may reduce the number of neuronal deaths and alleviate the sequelae of ischemic stroke. Meanwhile, our prior work demonstrated that elevated iron levels and neuronal iron accumulation in infarcted brain tissuecould be suppressed by deferiprone,[14] which mitigated neuronal death and reversed infarction-induced changes in key ferroptosis-related proteins (Gpx4 and Acsl4). Transmission electron microscopy further revealed ferroptosis-specific neuronal changes in the MCAO group, which were alleviated by deferiprone. These findings strongly indicate the presence of neuronal ferroptosis in ischemic stroke and further validate the reliability of the TOGGLE algorithm. Our work thus offers a new framework and toolset for investigating neuronal ferroptosis mechanisms and potential therapeutic targets. Moving forward, TOGGLE could be applied to map programmed cell death trajectories across various tissues and organs, supporting in-depth exploration of specific programmed cell death mechanisms.

Due to financial constraints and limitations of experimentation, the establishment of a stable and commercially accessible technology for in vivo cell fate prediction remains a challenge. Consequently, investigating the transcriptional diversity of cellular fate presents a difficulty. Nevertheless, our TOGGLE algorithm offers a computational approach to accomplish fate prediction, thereby significantly reducing the expenses associated with experimentation and facilitating research in developmental biology and cancer biology. Fate-specific changes in gene expression are accompanied by substantial epigenetic remodeling, which enhances the accessibility of the genome to fate-specifying transcription factors[6]. Therefore, the integration of scRNA-seq data and single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq) data to construct a multi-modality model is hypothesized to outperform a model solely based on scRNA-seq[40].

4 Methods

While there are numerous tools available for analyzing scRNA-seq data, few are designed to predict the fate of mature cells, such as neurons. Current computational methods include Weighted Optimal Transport (WOT)[34] and Cospar[7]. WOT constructs lineage information based on dynamic data trajectories but tends to misidentify similar transition trajectories. Cospar utilizes transition matrix principles to infer state transitions between nodes. However, it is not suitable for multi-stage classifications where correspondences are not unique; it cannot classify static, differentiated cell data,

such as Mature cells, nor can it distinguish between different programmed stages within the same cell type. The primary challenges in fate prediction, illustrated in **Figure 17**, include:

(1) Endpoint Loss. When using the above tools, information on both the starting point and endpoint of cell fate is required. For instance, it is necessary to have progenitor cells like Granulocyte-Monocyte Progenitors (GMP) that can give rise to progeny clones, and to capture samples of descendant clones such as Neutrophils and Monocytes. However, in certain diseases, it is impossible to obtain samples of cells that have reached their endpoint. For example, in ischemic stroke, neurons that have completed programmed cell death cannot be captured.

(2) Time Loss. The above algorithm requires sampling at multiple time points to piece together the continuous and gradual changes in cells, thereby reconstructing the trajectory of transition states during progenitor cell growth. However, in clinical practice, it is difficult to obtain samples at planned time intervals, such as post-surgical tumor tissues, hematomas after brain hemorrhage, or samples from deceased individuals or aborted embryos. Moreover, ethical and logistical concerns often repeated biopsies on humans at different time points. There is thus a compelling need for sensitive and accurate analytical approaches to predict cell fate from samples collected at a single time point. This requires first distinguishing cells at various stages of fate progression with high classification accuracy, followed by using these precise classifications to predict cell fate.

(3) Sample Imbalance. When collecting samples at multiple time points, it is possible to capture the starting point, endpoint, and intermediate processes of cell fate, along with dynamic information about the changes in proportion. Sample imbalance can be a significant issue not only at single timepoint but also across multiple timepoints. This problem arises due to the inherent limitations of single-cell technologies, such as uneven cell capture, sequencing depth variability, and the over- or underrepresentation of certain cell types. At a single time point, it is difficult to balance the number of cells in different stages of fate progression, leading to sparse data and uneven sample distribution. This usually ver-represent certain cell stages relative to the true distribution, because uneven cell numbers impacts both functional and type classification results. Cells with smaller populations and higher similarity are often misclassified by algorithms as the same type, this issue commonly arises in the identification of rare death types, complex immune functional subgroups, or rare subpopulations. This requires modifying the algorithm's principles, shifting from multi-time-point state transition construction to reconstructing continuous trajectory information from a single time point.

(4) Lack of Fate Boundaries. There is a distinct fate boundary between progenitor cells and Neutrophil/Monocyte cells, where transcriptional differences cause divergent fates. However,

ferroptosis and other forms of programmed cell death are often unidirectional and gradual fate progressions lacking distinct boundaries. Thus, cells undergoing programmed death/activat are often distributed across a spectrum of gene expression patterns that diverge from cells that are healthy, making the construction of the fate trajectory more challenging.

We named this method "Trajectory of Gene Generated Lineage Estimation (TOGGLE)." Based on the above issues, this study developed a new algorithm—TOGGLE. Compared with previous algorithms, TOGGLE offers the following advantages:

(1) This addresses a novel type of tracing problem. Except for immune cells, mature cells generally do not undergo cloning, making lineage tracing based on cell division unsuitable for mature cells. Mature cells exhibit highly similar temporal states, and their capture is often limited to a single stage within the programmed fate of cells.

(2) This method enables the tracing of mature cell fates or functions by constructing the continuous fate progression of mature cells (e.g., neurons) in highly similar datasets and grouping them according to their fate progression, even when clear fate boundaries are absent in the data.

(3) Unlike previous studies that rely solely on the cells at the starting point of a fate, TOGGLE can reconstruct the entire fate trajectory in complete datasets with unbalanced sample sizes, without requiring pre-identification of progenitor and descendant cells.

(4) A corollary of Takens' theorem is proposed, and the resulting method is termed modal entropy embedding. This represents a novel dynamic reconstruction theory based on chaos phase space reconstruction. Hence, this corollary is named the TOGGLE corollary.

(5) This algorithm not only establishes the developmental and differentiation lineage of cells (e.g., hematopoietic stem cells) and the reverse induction lineage of regeneration (e.g., reprogrammed cells), but also constructs the programmed fate lineage of mature cells (e.g., ferroptosis), functional transformations (e.g., cardiac pathway differences induced by e-cigarettes), functional lineages (e.g., cardiomyocytes), and functional subgroups (e.g., fibroblasts). TOGGLE thus enables the construction of the complete growth process of cells from development and differentiation to becoming mature cells. Furthermore, our novel tool facilitates the establishment of a comprehensive fate-tracking process for mature cells, including functional differences, functional transformations, and programmed cell death.

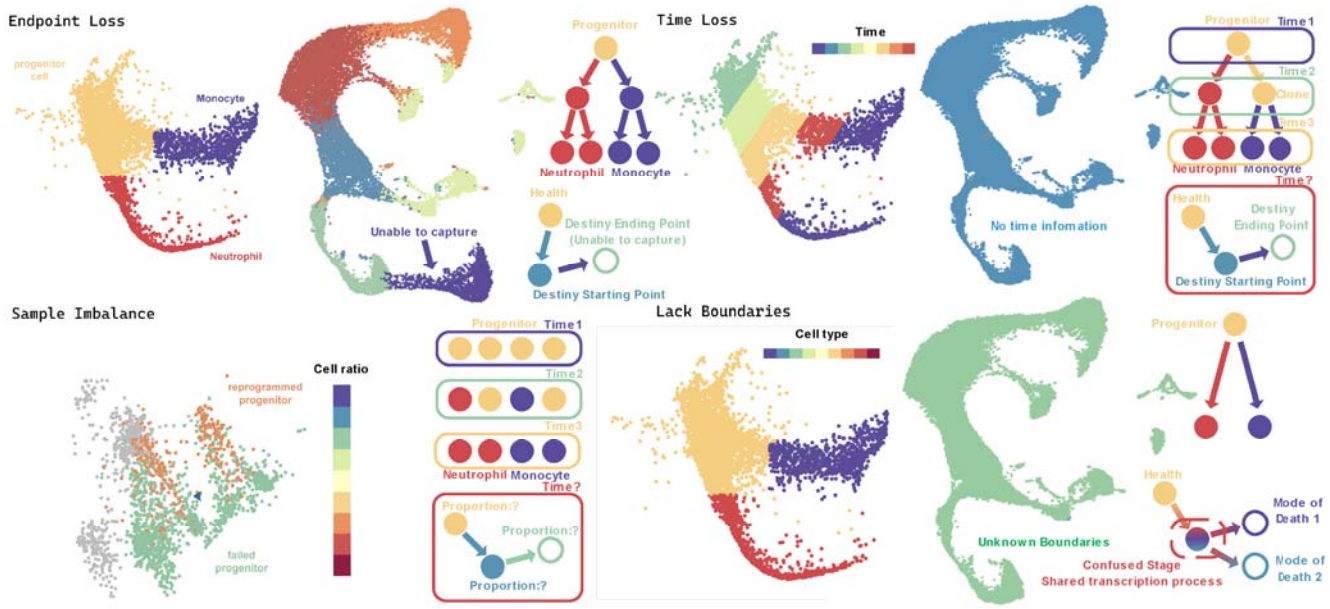

Figure 17. Challenges in Cell Fate prediction. Endpoint Loss: Current tools typically requires capturing cells at the endpoint of differentiation, whereas fate prediction may not capture this final state; Time Loss: Current tools require data from multiple time points, whereas fate prediction often relies on data from a single time point; Sample Imbalance: Fate prediction models frequently encounter imbalanced sample sizes; Lack of Boundaries: Previous tools depends on distinct boundaries between cell types, whereas fate prediction of neurons lacks clear boundaries between cell states. The sample data for the drawing are from GSE140802, GSE99915 and GSE232429.

4.1 Pipeline introduction

Self-supervised learning is a machine learning approach in which the algorithm leverages unlabeled data by constructing pretext tasks to generate reference hypotheses, which are then used for computation.The algorithm automatically assigns a presumed label by itself to the samples and these labels are used for learning. Consequently TOGGLE as an self-supervised algorithm. In initializing the inter-class metrics of cells, it is necessary to consider the phenomenon of metric center offset. When evaluating the similarity distribution S between cell A and surrounding cells with cell A as the reference center, one cannot infer that a similarity distribution like S exists when cell B is the reference center, meaning the similarity between cell A and cell B is typically not equal to the similarity between cell B and cell A. By modeling similarity metrics using the asymmetry of graph distances, it is possible to effectively represent this irreversibility of cell similarity distances. As shown in Figure 18, "1.2 Neighbor Activation", Input a gene expression data firstly, TOGGLE uses the graph diffusion in order to get the graph diffusion matrix(GDM) which has been proven effective successfully. After that,

TOGGLE uses learned features to keep the multiple kernels density consistency[41] , as illustrated in Figure 13, "Consistent Density by unsupervised Learning".

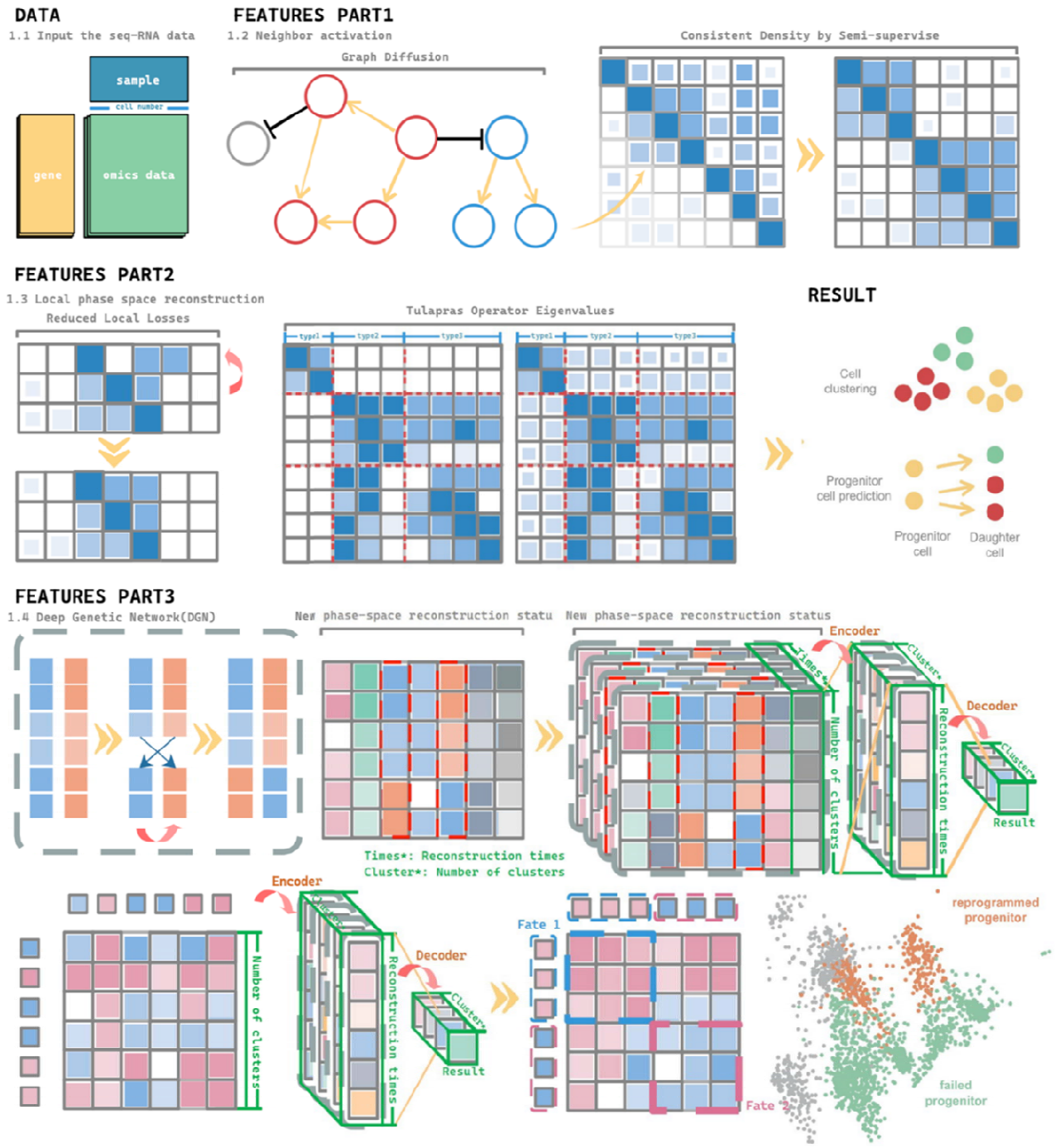

Figure 18 A data flow diagram illustrating the operational process will be provided, followed by a detailed analysis workflow diagram in later sections to aid biological readers in understanding the methodology. Part 1.1: Original high-dimensional data. Part 1.2, left of the yellow double arrow: Use Graph Diffusion to generate a distance matrix. Right of the yellow double arrow: Calculate correlation metrics and perform cell sorting. Part 1.3, left: Sort the compressed high-dimensional data to maximize the density of the effective region, a concept known as phase space reconstruction. Right: Obtain cells with similar transcriptional states. The distance matrix here can be used to identify similarities, also

referred to as a similarity matrix. Result: Clusters are obtained using a heuristic algorithm. Part 3, first row: Perform fitting with DGN, encode the results, and overlay them. Second row: Decode the encoded results to obtain the classification of clusters.

Following the completion of similarity measurement, it is necessary to perform sample sorting within clusters and boundary delineation. Unlike traditional algorithms that utilize absolute distances for sample sorting and defining inter-cluster boundaries, TOGGLE does not employ maximal eigenvalues or statistical features in multi-classification problems. Instead, it constructs relative distances using unsupervised principles and an iterative binarization hypothesis. At the start of the iteration, based on the singular value results after decentralization processing of the similarity matrix, the relative correlation index of all samples is calculated. Then, through the iterative process to reduce local losses, as shown in Reduced Local Losses in Figure 13, the binarization results are used as label hypotheses for unsupervised clustering. The initial clusters of this selfsupervised clustering are treated as the environmental response for unsupervised learning. By minimizing the loss function of environmental response, the hypothesis space internally re-differentiates relevance and repeats the iteration process. Using TOGGLE to complete the sorting of clusters within the similarity matrix achieves higher hierarchical clustering accuracy. As a result, cells will be clustered into subgroups, as well as different cells fate heterogeneity were clustered into subgroups with group ID.

In Figure 18-PART2, cells with similar states are clustered together, allowing several subgroups to be identified among progenitor cells, where cells within each subgroup share similar states. However, different subgroups may follow the same fate trajectory. For example, subgroups A and E could both represent early progenitor subpopulations, necessitating their grouping. In Figure 18-PART2, the sparse clustering of individual cells has been transformed into a non-sparse clustering of subgroups. We designed the Deep Genetic Network, an algorithm that maintains density consistency through phase space reconstruction. After each mutation using a genetic algorithm, a result is obtained, and each result is labeled as a code. By representing all possible results with these codes, the complete phase transition process in the feature space is reconstructed. Decoding the results clusters subgroups with similar states, such as grouping early progenitor subgroups A and E together. A heuristic algorithm is then used to determine the boundaries of the subgroups. Additionally, an extra test is conducted by mixing progenitor and progeny cells at different stages for classification, to demonstrate the algorithm's capability to classify mixed cells across categories.

Additionally, an extra test will be conducted by mixing progenitor and progeny cells at different stages for classification, in order to demonstrate the algorithm's ability to classify mixed cells across categories. To validate the performance of TOGGLE, we have applied it to two single-cell datasets:

Hematopoiesis and Fibroblast Reprogramming[8,9]. Upon analyzing the data, TOGGLE demonstrated its successful cell fate prediction capability and achieved state-of-the-art performance.

4.2 Use Graph Diffusion to generate cell distance

4.2 Graph Diffusion Activation Process

For the database, we used the expression data of scRNA-seq as the TOGGLE's input data. The datasets of scRNA-seq are from Gene Expression Omnibus database (accession No. GSE140802, GSE99915 and GSE99915)[1,6,42–45]. For the *simulated annealing* process for boundaries, we used built-in functions in MATLAB, because it's faster to deal with float and double data. All the data sets are available together with the open access database. Because the graph diffusion matrix exhibits asymmetry, the upper and lower halves of the matrix obtained by diagonal splitting contain different information. Moreover, graph diffusion utilizes inter-cellular graph distances for similarity measurement, which incorporates spatial distribution information. Each cell's transcriptional data affects the relative positioning of cells in the graph space, rather than relying on simple sequence alignments to determine similarity (Figure 19).

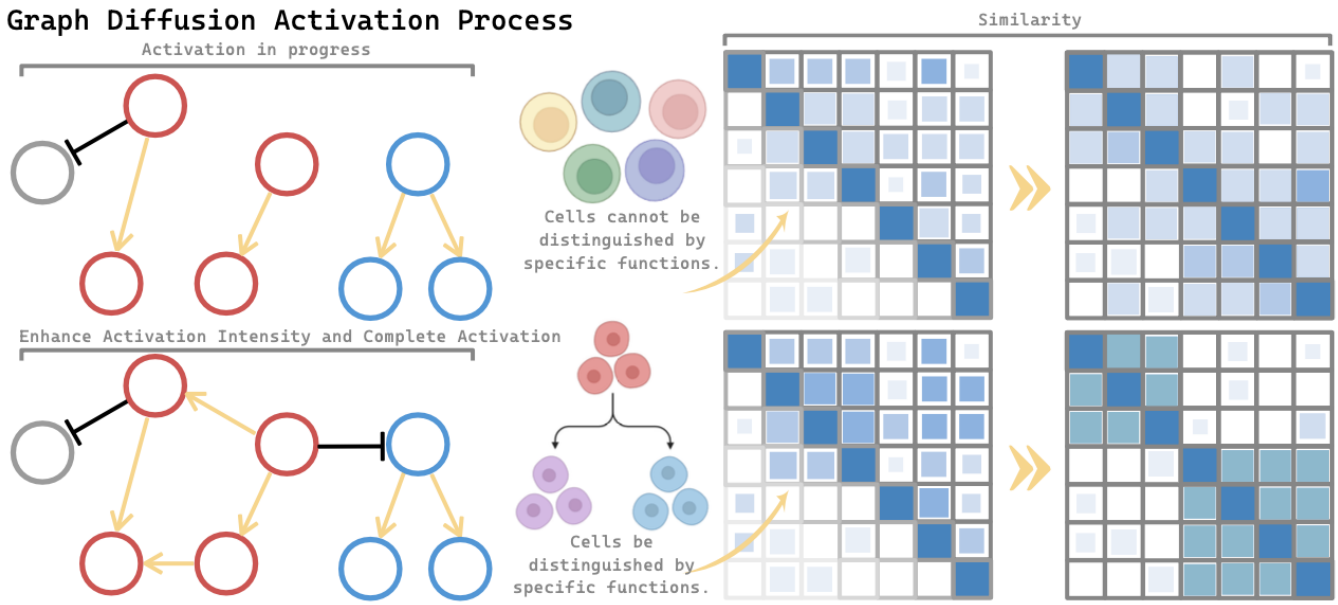

Figure 19 Generating functional similarities between cells using Graph Diffusion in deep learning. Top: In the initial Graph Diffusion activation, cell similarity is influenced by overall functional patterns. Bottom: In the improved approach, after completing full Graph Diffusion activation and performing dimensionality reduction and feature extraction, cell similarity is determined by cell-specific functions. Left: The Graph Diffusion activation process. Center: Differences in cell classification. Right: Cell-to-cell similarity mapping.

To obtain an RNA-seq expression matrix (M) of size $X \times Y$, transcriptional results for Y RNA molecules from X cells were retrieved from the NCBI database, resulting in a non-symmetric graph diffusion matrix of size $X \times X$. This matrix provides information about cell similarity. Extracting the $i-th$ row and $j-th$ column of the matrix yields M_{ij}, representing the one-way graph distance from the $i-th$ cell to the $j-th$ cell. Due to the asymmetry of the scatterplot distances, M_{ji} represents the one-way graph distance from the $j-th$ cell to the $i-th$ cell, typically not equal to M_{ij}.

Based on a common assumption that cells use the similarity matrix as a measure of class when cell populations are distributed around clustering centers denoted by d, clusters D are formed, centered around d. Ideally, the center d_i of cluster D_i is represented on the diagonal of the similarity matrix. Cluster D represents a highly similar group of cells within a subpopulation, generating label assumptions denoted as D'. TOGGLE utilizes these generated label assumptions D' for unsupervised learning. To generate predictive label assumptions D', it is necessary to use inter-cluster distances.

First, we construct a decentered matrix $\bar{D}_i$ for the similarity matrix M, allowing the data to be measured around the assumed virtual origin point.

$$\bar{D}_i = D_i - \frac{1}{n}\sum_{j=1}^{n} D_{ij}$$

Where D_i is the vector corresponding to the $i-th$ row of input matrix D, with $1 \leq i \leq m$. We compute the covariance matrix cov_D of $\bar{D}$ to enhance the dissimilarity in the similarity matrix, thereby mapping the original data to a latent space with greater differences.

$$cov_D = COV(\bar{D})$$

$COV(\bar{D})$ represents the covariance calculation matrix. Subsequently, we utilize the ideal feature matrix V, where the eigenvalues are distributed along the diagonal.

$$V = sparse((diag(cov_D))^{-\frac{1}{2}})$$

$diag(cov_D)$ denotes the extraction of diagonal elements from cov_D. The term "$sparse$" indicates the process of sparsifying the matrix. We achieve sparsification using "$sparse$" to obtain the sparsified diagonal matrix$(diag(cov_D))^{-\frac{1}{2}}$. Subsequently, we calculate the reconstruction matrix in the differential space to obtain the covariance similarity matrix pD.

$$pD = V \cdot cov_D \cdot V^T = \gamma(D)$$

Define this operation as $\gamma(D)$.Where pD represents the enhanced covariance matrix emphasizing dissimilarity. Using the definition of pD, we obtain the latent space mapping result $pM = \gamma\,(M)$ for

the similarity matrix , represented in the form of the correlation assumption ∂. represents the row vector in the matrix, with correlation coefficients for sample cells . Based on the numerical characteristics of the singular matrix, during the mapping process from the original object to the mapped object , cells with higher similarity in the original data will acquire closely related correlation coefficients. This effectively separates samples in accordance with the label assumptions, resulting in the formation of the sample set .

In simple terms, using amplifies the differences within the similarity matrix , making it easier to classify elements within the matrix.

4.3 Selfsupervised (A kind of Unspuervised) perform cell sorting

We proposed the TOGGLE corollary to demonstrate why this algorithm is effective. According to the corollary in Supplementary information-Section 5, using local phase space reconstruction allows for maximum consistency in information density. When applied to this biological problem, its intuitive significance lies in achieving maximum local similarity through the iterative reconstruction process. This ensures that cells in the most similar fate stages are aligned together, thus capturing the maximum amount of information (Figure 20).

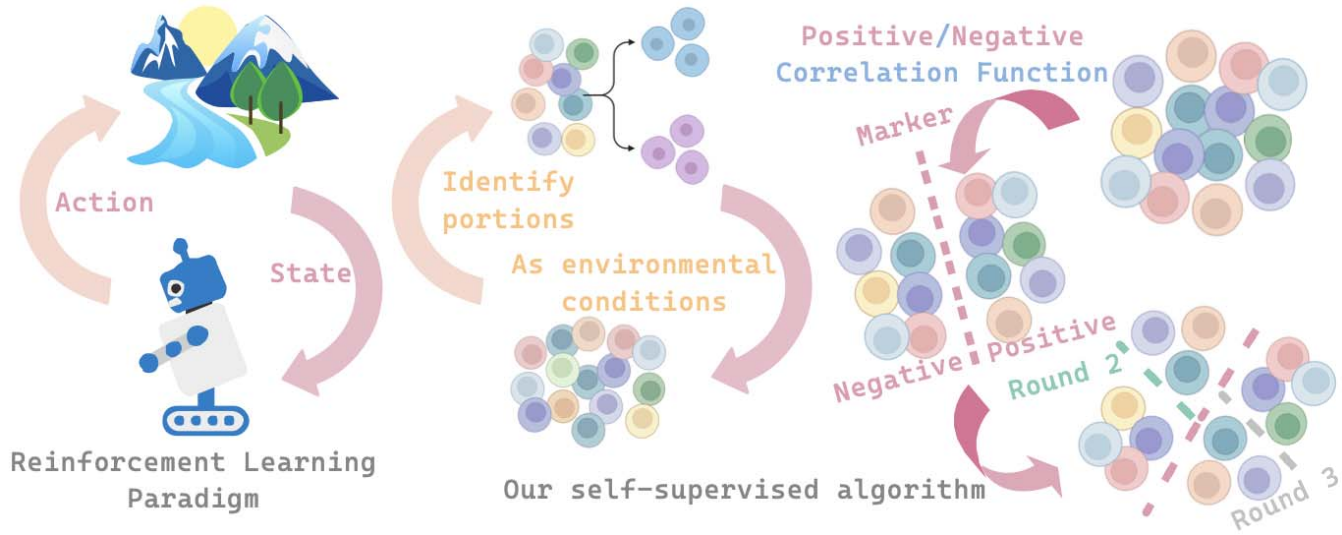

Figure 20 Methodological Schematic Diagram. The process begins by ranking cells and generating a boundary (Marker). The Marker serves as a condition to compare the correlation of each cell. Instead of using a traditional reward function, a correlation-based ranking loss index provides feedback, generating new Markers and refining the ranking iteratively. Cells positioned between two adjacent Markers are identified as a Cluster. Since Markers are autonomously generated by the algorithm, this method incorporates reinforcement learning feedback principles without requiring supervised training. The unsupervised framework leverages Marker-based self-feedback, eliminating the need for external labels.

(Left: Reinforcement learning principles; Center: Theoretical modeling and imitation; Right: Practical functional clustering implementation)

The detailed explanation of $loss(\partial_n)$ can be found in Supplementary information-Section 6. When directly using correlation coefficients for classification, there is a presence of numerous boundary samples ∂_z between different clusters, which makes the boundaries less distinct or prone to being misclassified as independent clusters. Assuming that ∂_d is located at the clustering center of cluster D, when the convergence of cluster D within its hypothetical space Sp occurs, there exists a loss function, denoted as $loss(\partial_n)$:

$$loss(\partial_n) = \lim_{n \to d} |\partial_n - \mathrm{d}_n| < \varepsilon$$

Where ε represents the theoretical threshold range. Here, d_n represents samples with the same label within cluster D. Since an ideal cluster D represents a functionally homogeneous group of the same cell type, in an ideal cluster D, it holds that $D\{d_1 = d_2 = \ldots = d_n\}$. When applying binaryization assumptions to the samples using the parameter δ, the mapping process of any sample $\partial_1 \sim \partial_z$ in the space Sp, as defined in the label assumption set D', is as follows:

$$[\partial_1 \sim \partial_z, \partial_\pi] \rightarrow [\delta_1, \delta_2, \delta_3, \ldots, \delta_{n-1}, \delta_n] = Sp$$

Where ∂_π represents non-boundary samples, which do not fall within the theoretically defined boundary vicinity interval $(\partial_\pi - \varepsilon, \partial_\pi + \varepsilon)$. In this scenario, the minimization of the loss function, denoted as $loss(\delta_n)$, is achieved. Based on the convergence properties of the loss function $loss(\delta_n)$, it is possible to minimize the inversion number of the space matrix Sp, denoted as $\tau(\delta_1, \delta_2, \delta_3, \ldots, \delta_{n-1}, \delta_n)$. Starting from a correlation origin point of 0, clusters can be represented in the minimized form of the inversion number τ as D^+ and D^-:

$$D^+ = \{0, \partial^+{}_n, \partial^+{}_n, \ldots, \partial^+{}_{n+\alpha}\}$$

$$D^- = \{\partial^-{}_m, \partial^-{}_m, \ldots, \partial^-{}_{m+\beta}, 0\}$$

Where $\partial^+{}_n$ represents samples for which the binary assumption value δ_n is positive, denoted as $\partial^+{}_n$, and $\partial^-{}_m$ m represents samples for which the binary assumption value δ_m is negative, denoted as $\partial^-{}_m$. Subsequently, TOGGLE can utilize the self-labeled binary assumption labels D^+ and D^- to complete the Markov decision process. The entire process can be represented as $\{D^+, D^-\} = g(D)$, when D satisfies the conditions of environmental response determination:

$$agent(D) = \exists\{\mathrm{t}\,[g(D)] \neq 0\}$$

Once the conditions for unsupervised learning are met, the reward mapping $D \rightarrow [D^+, D^-]$ is provided to indicate environmental responses. Similarly, when D^- satisfies the condition determined by $agent(D^-)$, a D^- branch is initiated as $\{D'^+, D'^-\} = g(D^-)$, until all samples in D^- are represented as a single mapping $D^- \rightarrow [D'^-]$. Each mapping process only adheres to the current state dictated by the environment response, ensuring that each iteration possesses a consistent invariance. It solely relies on the current hypothesis space Sp and is independent of the order of samples. In other words, during iterations, randomly shuffling a few samples does not affect the iteration results. Whenever a new branch is generated, the decision-making process $agent(D)$ depends on the inversion number $agent(D', D'', D''', \ldots)$ encompassing all mapping processes:

$$agent(D', D'', D''', \ldots) = \exists\{\tau[g(D)] \neq 0, \tau[g(\{D^+)] \neq 0, \tau[g(D^-)] \neq 0, \ldots\}$$

During each iteration of the representation process $\{D^+, D^-\} = g(D)$, boundaries in the state of the Markov chain are generated. These boundaries, denoted as s, allow us to determine the boundaries of cluster D. As the iterations progress, sample boundaries within D^+ and D^- can be obtained, denoted as s'. The number of boundaries required can be selected based on the resolution requirements.

In simple terms, the covariance similarity matrix pD exhibits square patterns along its diagonal, where the inside of each square has higher similarity values, defining what we refer to as cluster D. We employ the loss function $loss(\partial_n)$ to assess the convergence of this cluster. Each cluster D has a central point ∂_d with the highest similarity, which corresponds to its true label d_n. When a cell sample ∂_n is assigned to cluster D, the cell's assumed label is also marked as d_n. However, ∂_d may have multiple potential labels, making it challenging to achieve uniformity within cluster D and potentially leading to the subdivision of cluster D into various smaller categories, making it difficult to measure the loss function value.

However, when ∂_n is transformed into binary values, i.e., solely discerning whether the sample ∂_n is positively correlated with the cluster's core ∂_d or not, if positively correlated, it is assigned to cluster D^+; otherwise, it goes to D^-. After a single iteration, the results of the iteration, D^+ and D^-, are assumed to be the closest to the true labels. Based on these results, the convergence of samples within D^+ and D^- is continuously evaluated. Taking D^+ as an example, if not all samples within D^+ are positively correlated with the center of D^+, D^+ may be further subdivided into D'^+ and D'^- based on their correlations, with D' representing a new iteration of D. Positive correlation between samples and cluster centers indicates convergence, while the presence of non-converging samples necessitates the

delineation of boundaries between converging and non-converging samples. When computing samples within , the knowledge reservoir used consists of and , which have already been transformed into unsupervised labels from test samples. When the entire matrix converges completely during the iterations, the sorting and segmentation process of the matrix is considered complete.

4.4 Unsupervised Deep Genetic Network

Using correlation makes it easy to determine the relative correlation distance of a cell with respect to the Marker. By selecting cells located between adjacent Markers, we can identify functionally similar cells. However, the relationships between cellular functions are often nonlinear, meaning that different cell types with distinct functions may still exhibit similar functional relationships. Thus, correlation alone is insufficient to determine the final cell type classification. To address this, we developed a Deep Genetic Network, adapting the Mask framework from BERT to an unsupervised setting. This approach unifies the global similarity relationships between a single cell group and all other groups, enabling precise functional classification of Clusters. The final functional assignment of each cluster is determined through this framework, and its theoretical validity is rigorously analyzed using the inference of Takens' theorem. The detailed principles of this approach are illustrated in the corresponding figure 21.

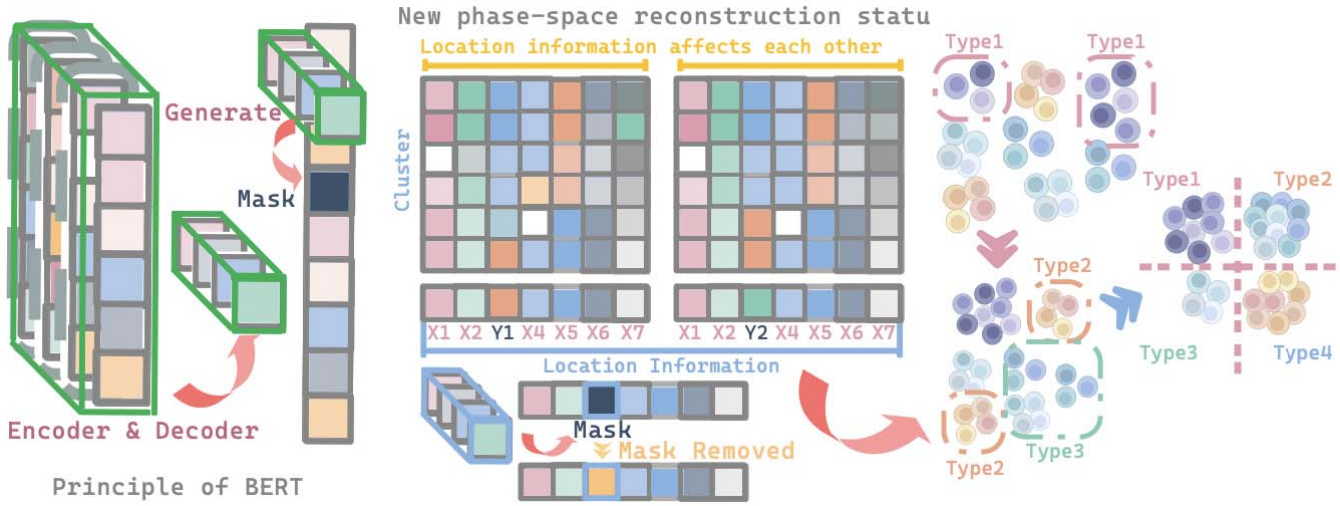

Figure 21 Schematic Diagram of the Methodology Principles. Left: Inspired by BERT's MASK mechanism, the Mask framework is introduced into an unsupervised learning setting. Since genetic algorithms use genetic variation instead of parameter fitting, this framework does not require pre-training. Center: The positioning of a given cluster is determined by its inter-cluster distance relative to others, meaning that there is a contextual Marker semantic relationship (global information). However, self-supervised Markers are not always entirely accurate, as they ignore global information and rely on linear inference. The genetic algorithm introduces random mutations, recalculating inter-cluster distances and fitting results, effectively refining global information. When a mutation produces a new outcome, the existing position information is removed and redefined using the Mask mechanism. Right:

Intuitive representation of the algorithm's function, demonstrating the process of recalculating inter-cluster distances.

After completing the calculations in the previous section, the sorted matrix pM is obtained. The internal boundary of pM is $S = \{s', s'', s'''\ldots S^{n\prime}\}$, where $S^{n\prime}$ is the boundary generated during the n-th iteration. S is the predictive information obtained through unsupervised learning. After using S to segment pM, a differentiation matrix pM for cell fate stages can be obtained. At this point, the cell similarity matrix pM is transformed into a similarity matrix rM for fate stages. According to the Corollary of Takens' theorem in Supplementary information-Section 5, setting $Y = rM$ allows the conclusion $E(Y_i, \varsigma) \Leftrightarrow E(Y_i)$ to be applied.

About Supplementary information-Section 5, simply put, the results generated by the genetic algorithm have a certain randomness. After sufficiently iterating the algorithm, the complete distribution of results under fixed conditions can be obtainedence the spatial distribution of conditions is sufficiently represented, the distribution of results will remain stable. The maximum distribution within the result distribution is taken as the final classification result, which proves the validity of the conclusion $E(Y_i, \varsigma) \Leftrightarrow E(Y_i)$, as the maximum information density is achieved at this point.

Define the genetic algorithm as $Genetic$. In this context, a single result produced by the genetic algorithm$Genetic(rM)$ is an observation of the true result under specific conditions ς_i , denoted as $E(rM_i, \varsigma_i)$. $E(rM)$ is the ideal true fitted result. By encoding $E(rM_i, \varsigma_i)$, we can obtain $Ec[E(rM_i, \varsigma_i)] = R_i$. $Count(R_i, \varsigma)$ represents the count of occurrences of R_i under the condition set ς. Decoding the result yields

$$Ed(rM, \varsigma) = arg \underbrace{max}_{\varsigma i \in \varsigma} Count(R_i)$$

Here Ec represents encoding, Ed represents decoding, and R_i represents a type of fitted result. For visualization purposes, we employ dimensionality reduction algorithms to project data into two or three dimensions, allowing us to intuitively depict the hidden structures within the data.

Subsequently, a heuristic algorithm is used to perform boundary partitioning for $Ed(rM, \varsigma)$. The valid sampling range for $Ed(rM, \varsigma)$ is $Ed(rM, \varsigma) \in (\alpha, \beta)$, with the sampling resolution P. Here set $\beta = \alpha + \gamma P$, resulting in a Bernoulli coding matrix, i.e., the valid information matrix:

$$Ed(rM, \varsigma)_{Boolean} = \begin{cases} Ed(rM, \varsigma) \in (\alpha, \alpha + iP) = 1 \\ Ed(rM, \varsigma) \notin (\alpha, \alpha + iP) = 0 \end{cases} \quad i \in (\alpha, \alpha + \gamma P)$$

The encoded result of $Ed(rM, \varsigma)$ is $E_{code} = \sum_{i=0}^{\gamma P} Ed(rM, \varsigma)_{Boolean}$, Decoding E_{code} yields the boundary $Decode\left[\sum_{i=0}^{\gamma P} Ed(rM, \varsigma)_{Boolean}\right]$, defined as follows:

$$Decode = \begin{cases} E_{code}(j,k) \neq E_{code}(j,k+1) \rightarrow Decode(j,k+1) = Decode(j,k+1) + 1 \\ E_{code}(j,k) \neq E_{code}(j+1,k) \rightarrow Decode(j+1,k) = Decode(j+1,k) + 1 \end{cases}$$

Using $Decode$, the frequency of boundary occurrences can be determined. Simulated annealing optimization is then applied to obtain the boundary. Using the same approach(Sledge & Príncipe, 2018), resulting in an $N \times B$ latent matrix Z, and assign hierarchical clustering labels to each cell. We employ a heuristic approach based on graph Laplacian decomposition to estimate the number of clusters C, with the reduced dimensionality B being equal to the desired number of clusters C, where C is also the aforementioned ranking-constrained parameter.

4.5 scRNA-seq Data Processing

Cells from the GSE232429 dataset[1] with nFeature_RNA > 200 and < 7000, nCount_RNA > 500 and < 25000, and percent.mt < 10% were retained for further analysis. DEsingle was used to perform a preliminary analysis of this dataset to identify differentially expressed genes. Differentially expressed genes were screened out using the criteria of |Log2FC|≥0.25 and FDR<0.05. Cellchat was used for cell-cell communication analysis[46].

4.6 Experimental Materials and Methods

4.6.1 Experimental Materials

(1) Experimental Animal

Fifty male specific pathogen free Sprague-Dawley (SD) rats, aged 6 to 8 weeks and weighing 250-280 g, were provided by the Experimental Animal Center of Hunan University of Chinese Medicine (Animal Quality Certificate No.: 430727211100405318). The rats were purchased from Hunan Sileike Jingda Experimental Animal Co., Ltd. [License No.: SCXK (Xiang) 2019-0004]. They were housed in the Experimental Animal Center of Hunan University of Chinese Medicine [License No.: SYXK (Xiang) 2019-0009] under standard conditions, with a room temperature of (24±1)°C, humidity of approximately 55%, and a 12-hour light/dark cycle. After a one-week acclimation period, the animals were used for the experiments. All animals' care and experimental procedures were approved by the Animal Ethics Committee of Hunan University of Chinese Medicine (Approve number: LL2019092009) and were in accordance with the National Institute of Health's Guide for the Care and Use of Laboratory Animals.

(2) Reagents

Paraformaldehyde (Batch No.: G1101) was purchased from Servicebio, and anhydrous ethanol (Batch No.: 10009259) from China National Pharmaceutical Group Chemical Reagent Co., Ltd. Electron microscopy fixative (Cat. No. G1102) was purchased from Wuhan Guge Biotechnology Co., Ltd. Transferrin (Trf) polyclonal antibody (Cat. No.: 17435-1-AP), Bax polyclonal antibody (Cat. No.: 50599-2-Ig) and IκBα polyclonal antibody (Cat. No.: 10268-1-AP) were obtained from Proteintech.

Casp-3 antibody (Cat. No.: 9661L) was purchased from Cell Signaling. Sirt1 antibody (Cat. No. ab189494) was purchased from Abcam. Transferrin receptor (Tfrc) polyclonal antibody (Cat. No.: A5865), and Rela (p65) polyclonal antibody (Cat. No.: A2547) were sourced from ABclonal. P-smad (Cat. No.: AF8313) was acquired from Affinity Bioscience. Gpx4 antibody (Cat. No.: AF301385), Acsl4 polyclonal antibody (Cat. No.: AFW6826) and Fsp1 polyclonal antibody (Cat. No.: AFW21808) were purchased from AiFangbio. DAPI (Cat. No.: ab228549) was obtained from Abcam. Deferiprone tablets (specification: 50 mg/tablet, batch number: H20140379) was purchased form Apotex Inc.

(3) Instruments

The optical microscope (Axioscope-5) was purchased from Zeiss, Germany. The pathology microtome (RM2016) and Leica blades (LEICA-819) were both obtained from Leica Instruments Co., Ltd. Transmission electron microscope (HT7700) was purchased from Hitachi Inc.

4.6.2 Animal Grouping, Modeling and Intervention

Fifty SD rats were randomly assigned into three groups: the Sham operation group, the MCAO group and the MCAO+Deferiprone group, with five rats in each group. The MCAO model was established using a modified reversible middle cerebral artery suture method[47] for the rats in the MCAO group and the MCAO+Deferiprone group. Rats were fasted for 12 hours prior to surgery. During surgery, the right common carotid artery and external carotid artery were exposed and clamped, and a suture was inserted into the internal carotid artery via the common carotid artery to block the origin of the right middle cerebral artery. The suture was advanced approximately 18.5±0.5 mm until resistance was felt, indicating the correct position. The suture was then fixed to both the common and internal carotid arteries, and the incision was closed. Throughout the procedure, care was taken to maintain body temperature, and after surgery, the animals were returned to normal feeding conditions. In the Sham group, no suture was inserted, and the animals were kept under normal conditions.

Successful model establishment was confirmed when the rats exhibited left-sided limb paralysis, instability when standing, or circling behavior when lifted by the tail. Neurological function was scored on a scale of 1 to 3, with non-successful models excluded. The neurological scoring system was based on the Zea Longa scale[47] as shown in Table S3.

Deferiprone tablets were dissolved in a 0.9% sodium chloride solution. Following model establishment, rats in the deferiprone group received oral administration of deferiprone at a dose of 125 mg/kg per day. Rats in the sham operation group and the MCAO group were given an equivalent volume of 0.9% sodium chloride solution. Treatments were administered once daily for three consecutive days.

4.6.3 Neuronal morphological changes observation by hematoxylin-eosin (HE) staining

The brain tissue was perfused with pre-cooled 0.9% sodium chloride solution, followed by fixation in 4% paraformaldehyde at 4°C. The tissue was then coronally sectioned into 4 μm thick slices. The sections were deparaffinized, rehydrated, and stained with hematoxylin for 5 minutes. After differentiation and rinsing with water, the sections were blued with distilled water, dehydrated through a graded ethanol series, and counterstained with eosin for 5 minutes. Following dehydration with ethanol and clearing with xylene, coverslips were applied and sealed with neutral balsam. Morphological observations and image acquisition were performed under a light microscope.

4.6.4 Transmission electron microscopy

The affected brain tissue was sectioned into 1 mm-thick slices. Samples were immersed in electron microscopy fixative at 4°C overnight, followed by fixation in 1% osmium tetroxide at 4°C for 2 hours. The samples were stained with 1% uranyl acetate for 1 hour, dehydrated, and embedded in epoxy resin. Ultrathin sections were prepared using an ultramicrotome, stained with uranyl acetate and lead citrate, and examined under a transmission electron microscope.

4.6.5 Immunohistochemical Staining for Acsl4, Gpx4 and Fsp1 Protein Expression Levels

Paraffin-embedded tissue sections (5 μm) were deparaffinized and rehydrated. Endogenous enzymes were inactivated by incubating the sections in 3% hydrogen peroxide at room temperature for 10 minutes, followed by two washes with distilled water. Antigen retrieval was performed by immersing the sections in 0.01 mol/L citrate buffer (pH 6.0) and heating in a microwave until boiling, then turning off the power and repeating the process after a 5-minute interval. After cooling, the sections were washed twice with PBS buffer (pH 7.2-7.6) for 3 minutes each. Blocking solution was added and incubated at room temperature for 20 minutes, with excess liquid removed without washing. Primary antibodies were applied and incubated at 37°C for 2 hours, followed by three washes with PBS buffer (pH 7.2-7.6) for 3 minutes each. Biotinylated secondary antibody solution was added and incubated at 20-37°C for 20 minutes, followed by another round of PBS washing. Streptavidin-HRP was applied and incubated at 20-37°C for 20 minutes, followed by washing in PBS as described earlier. DAB was used for color development, with reaction time controlled under a microscope (5-30 minutes). Sections were lightly counterstained with hematoxylin, dehydrated, cleared, and mounted with neutral balsam for microscopic observation of Acsl4 and Gpx4 protein expression. Images from five randomly selected fields of view from different samples were captured, and average optical density was analyzed using ImageJ software.

4.6.6 Immunofluorescence Staining for Bax, IκBα, Casp3, p-Smad, p65 (Rela), Tfrc, Trf, and Sirt1 Protein Expression Levels

Paraffin sections were deparaffinized and rehydrated, followed by membrane permeabilization with 0.3% Triton X-100 for 5 minutes. The sections were then blocked with 5% BSA in PBST at room temperature for 1 hour. Primary antibodies, diluted in PBST, were applied, and the sections were incubated overnight at 4°C. After washing with PBST, fluorescent secondary antibodies were added, and the sections were incubated in the dark at room temperature for 1 hour. The secondary antibodies were washed off, and the slides were mounted with anti-fade mounting medium containing DAPI. The next day, images were captured using a panoramic tissue cell quantification system at 40x magnification. Images from five randomly selected fields of view from different samples were captured and analyzed using ImageJ software to assess the average fluorescence intensity.

4.7 Statistical Analysis

All data are presented as Mean±SD, and statistical analysis and graphing were performed using GraphPad Prism 9.0. Non-parametric tests were used to analyze average fluorescence intensity and average optical density. A P-value of less than 0.05 was considered statistically significant.

Code and Data availability

The raw data for the hematopoiesis dataset can be accessed at the Gene Expression Omnibus database with accession number GSE140802, the reprogramming dataset with accession number GSE99915, the e-cigarette aerosol inhalation on mouse cardiac tissue with accession number GSE183614, the rat myocardial infarction dataset with accession number GSE253768, the ischemic stroke dataset with accession number GSE232429, and the adult NSC lineage dataset with accession number GSE209656 and GSE211786. All codes and analyzed data are available at:

https://github.com/FullBlackWolf/TOGGLE?tab=readme-ov-file

https://fullblackwolf.github.io/TOGGLE/

Contribution

The methodology, bioanalysis, and test application of this paper were collaboratively designed by Junpeng Chen*+, Eric Jiang+, Zhouweiyu Chen+, and Kailin Yang+*. Junpeng Chen organized and led the entire study, defining the manuscript's topic and narrative. Junpeng Chen and Zhouweiyu Chen were the main contributors to the algorithm innovation, while Eric Jiang and Kailin Yang were the main contributors to the bioanalysis and bio-innovation. Zhouweiyu Chen assisted in solving critical computational problems and completed the design and testing of auxiliary plans, making the method complete. Eric Jiang proposed a complete set of biological problems and performed the analysis, with assistance from Junpeng Chen, Kailin Yang, Kaiqing Liu, Tianda Sun, Yibing Nong, Tao Yuan, Pawan

Daga, Charles C. Dai, Changping Li, Yexing Yan, Zixiang Su, You Li, Tian Song and Richa A. Singhal. Jinwen Ge, Haihui Wu, Tong Yang, Shanshan Wang and Kailin Yang performed the animal experiments. Alex P. Carll played a crucial role in correspondence, providing experimental resources and research funding. Junpeng Chen, Alex P. Carll, and Lu Cai jointly organized the study. Alex P. Carll, Yibing Nong guided the analysis of cardiac pathway differences and cardiomyocytes. Lu Cai supervised the experimental work on ferroptosis.

Acknowledgments

This work was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health and the United States Food and Drug Administration Center for Tobacco Products (R01HL163818 to APC). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or FDA. "***This research was supported in part by the U.S. National Science Foundation (NSF) under grant CNS***1828521 ***and the University of Louisville's Research Computing team***." *The University of Louisville provided computational equipment support for this research and assisted all researchers with equipment maintenance and updates. These hardworking and dedicated staff members share this honor with us, including the following staff*: ***Harrison Simrall***, ***Tomas Felipe Llano Rios***, and ***Randy Vargas***.

References

1. Nakamura, A., Sakai, S., Taketomi, Y., Tsuyama, J., Miki, Y., Hara, Y., Arai, N., Sugiura, Y., Kawaji, H., Murakami, M. & Shichita, T. PLA2G2E-mediated lipid metabolism triggers brain-autonomous neural repair after ischemic stroke. *Neuron* **111,** 2995-3010.e9 (2023).
2. Kremer, L. P. M., Cerrizuela, S., El-Sammak, H., Al Shukairi, M. E., Ellinger, T., Straub, J., Korkmaz, A., Volk, K., Brunken, J., Kleber, S., Anders, S. & Martin-Villalba, A. DNA methylation controls stemness of astrocytes in health and ischaemia. *Nature* **634,** 415–423 (2024).
3. Tuo, Q.-Z., Zhang, S.-T. & Lei, P. Mechanisms of neuronal cell death in ischemic stroke and their therapeutic implications. *Med Res Rev* **42,** 259–305 (2022).
4. Li, M., Meng, Z., Yu, S., Li, J., Wang, Y., Yang, W. & Wu, H. Baicalein ameliorates cerebral ischemia-reperfusion injury by inhibiting ferroptosis via regulating GPX4/ACSL4/ACSL3 axis. *Chem Biol Interact* **366,** 110137 (2022).
5. Yang, K., Zeng, L., Yuan, X., Wang, S., Ge, A., Xu, H., Zeng, J. & Ge, J. The mechanism of ferroptosis regulating oxidative stress in ischemic stroke and the regulation mechanism of natural pharmacological active components. *Biomedicine & Pharmacotherapy* **154,** 113611 (2022).
6. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367,** eaaw3381 (2020).

7. Wang, S.-W., Herriges, M. J., Hurley, K., Kotton, D. N. & Klein, A. M. CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nat Biotechnol* **40,** 1066–1074 (2022).
8. Moreno-Ayala, R. & Junker, J. P. Single-cell genomics to study developmental cell fate decisions in zebrafish. *Brief Funct Genomics* elab018 (2021). doi:10.1093/bfgp/elab018
9. Biddy, B. A., Kong, W., Kamimoto, K., Guo, C., Waye, S. E., Sun, T. & Morris, S. A. Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564,** 219–224 (2018).
10. Neczypor, E. W., Saldaña, T. A., Mears, M. J., Aslaner, D. M., Escobar, Y.-N. H., Gorr, M. W. & Wold, L. E. e-Cigarette Aerosol Reduces Left Ventricular Function in Adolescent Mice. *Circulation* **145,** 868–870 (2022).
11. Qin, H., Abulaiti, A., Maimaiti, A., Abulaiti, Z., Fan, G., Aili, Y., Ji, W., Wang, Z. & Wang, Y. Integrated machine learning survival framework develops a prognostic model based on inter-crosstalk definition of mitochondrial function and cell death patterns in a large multicenter cohort for lower-grade glioma. *Journal of Translational Medicine* **21,** 588 (2023).
12. Zhou, N., Yuan, X., Du, Q., Zhang, Z., Shi, X., Bao, J., Ning, Y. & Peng, L. FerrDb V2: update of the manually curated database of ferroptosis regulators and ferroptosis-disease associations. *Nucleic Acids Research* **51,** D571–D582 (2023).
13. Kanehisa, M., Sato, Y. & Kawashima, M. KEGG mapping tools for uncovering hidden features in biological data. *Protein Sci* **31,** 47–53 (2022).
14. Lan, B., Ge, J.-W., Cheng, S.-W., Zheng, X.-L., Liao, J., He, C., Rao, Z.-Q. & Wang, G.-Z. Extract of Naotaifang, a compound Chinese herbal medicine, protects neuron ferroptosis induced by acute cerebral ischemia in rats. *J Integr Med* **18,** 344–350 (2020).
15. Younesi, F. S., Miller, A. E., Barker, T. H., Rossi, F. M. V. & Hinz, B. Fibroblast and myofibroblast activation in normal tissue repair and fibrosis. *Nat Rev Mol Cell Biol* **25,** 617–638 (2024).
16. Alvarez, S. W., Sviderskiy, V. O., Terzi, E. M., Papagiannakopoulos, T., Moreira, A. L., Adams, S., Sabatini, D. M., Birsoy, K. & Possemato, R. NFS1 undergoes positive selection in lung tumours and protects cells from ferroptosis. *Nature* **551,** 639–643 (2017).
17. Ngo, M. & Ridgway, N. D. Oxysterol Binding Protein–related Protein 9 (ORP9) Is a Cholesterol Transfer Protein That Regulates Golgi Structure and Function. *MBoC* **20,** 1388–1399 (2009).
18. Polat, I. H., Tarrado-Castellarnau, M., Bharat, R., Perarnau, J., Benito, A., Cortés, R., Sabatier, P. & Cascante, M. Oxidative Pentose Phosphate Pathway Enzyme 6-Phosphogluconate Dehydrogenase Plays a Key Role in Breast Cancer Metabolism. *Biology (Basel)* **10,** 85 (2021).
19. Yu, Y., Zhang, M., Liu, J., Xu, B., Yang, J., Wang, N., Yan, S., Wang, F., He, X., Ji, G., Li, Q. & Miao, L. Long Non-coding RNA PVT1 Promotes Cell Proliferation and Migration by Silencing ANGPTL4 Expression in Cholangiocarcinoma. *Molecular Therapy - Nucleic Acids* **13,** 503–513 (2018).
20. Blankman, J. L., Long, J. Z., Trauger, S. A., Siuzdak, G. & Cravatt, B. F. ABHD12 controls brain lysophosphatidylserine pathways that are deregulated in a murine model of the neurodegenerative disease PHARC. *Proc Natl Acad Sci U S A* **110,** 1500–1505 (2013).
21. Rashid, M. A., Lee, S., Tak, E., Lee, J., Choi, T. G., Lee, J.-W., Kim, J. B., Youn, J. H., Kang, I., Ha, J. & Kim, S. S. Carbonyl reductase 1 protects pancreatic β-cells against oxidative stress-induced apoptosis in glucotoxicity and glucolipotoxicity. *Free Radical Biology and Medicine* **49,** 1522–1533 (2010).
22. Tang, D. & Kroemer, G. Peroxisome: the new player in ferroptosis. *Sig Transduct Target Ther* **5,** 273 (2020).
23. Ou, Y., Wang, S.-J., Li, D., Chu, B. & Gu, W. Activation of *SAT1* engages polyamine metabolism with p53-mediated ferroptotic responses. *Proc. Natl. Acad. Sci. U.S.A.* **113,** (2016).
24. Scarpellini, G., Fiorini, S., Giuliari, F., Moreiro, P. & Del Bue, A. Diffassemble: A unified graph-diffusion model for 2d and 3d reassembly. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 28098–28108 (2024).

25. Richter, T., Bahrami, M., Xia, Y., Fischer, D. S. & Theis, F. J. Delineating the effective use of self-supervised learning in single-cell genomics. *Nat Mach Intell* **7,** 68–78 (2024).
26. Linder, J., La Fleur, A., Chen, Z., Ljubetič, A., Baker, D., Kannan, S. & Seelig, G. Interpreting neural networks for biological sequences by learning stochastic masks. *Nat Mach Intell* **4,** 41–54 (2022).
27. Raghavan, G., Tharwat, B., Hari, S. N., Satani, D., Liu, R. & Thomson, M. Engineering flexible machine learning systems by traversing functionally invariant paths. *Nat Mach Intell* **6,** 1179–1196 (2024).
28. Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H. B., Pe'er, D. & Theis, F. J. CellRank for directed single-cell fate mapping. *Nat Methods* **19,** 159–170 (2022).
29. Bannon, D., Moen, E., Schwartz, M., Borba, E., Kudo, T., Greenwald, N., Vijayakumar, V., Chang, B., Pao, E., Osterman, E., Graf, W. & Van Valen, D. DeepCell Kiosk: scaling deep learning–enabled cellular image analysis with Kubernetes. *Nat Methods* **18,** 43–45 (2021).
30. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat Methods* **16,** 715–721 (2019).
31. Domínguez Conde, C., Xu, C., Jarvis, L. B., Rainbow, D. B., Wells, S. B., Gomes, T., Howlett, S. K., Suchanek, O., Polanski, K., King, H. W., Mamanova, L., Huang, N., Szabo, P. A., Richardson, L., Bolt, L., Fasouli, E. S., Mahbubani, K. T., Prete, M., Tuck, L., Richoz, N., Tuong, Z. K., Campos, L., Mousa, H. S., Needham, E. J., Pritchard, S., Li, T., Elmentaite, R., Park, J., Rahmani, E., Chen, D., Menon, D. K., Bayraktar, O. A., James, L. K., Meyer, K. B., Yosef, N., Clatworthy, M. R., Sims, P. A., Farber, D. L., Saeb-Parsy, K., Jones, J. L. & Teichmann, S. A. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376,** eabl5197 (2022).
32. Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., Van Den Oord, J., Atak, Z. K., Wouters, J. & Aerts, S. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14,** 1083–1086 (2017).
33. Van De Sande, B., Flerin, C., Davie, K., De Waegeneer, M., Hulselmans, G., Aibar, S., Seurinck, R., Saelens, W., Cannoodt, R., Rouchon, Q., Verbeiren, T., De Maeyer, D., Reumers, J., Saeys, Y. & Aerts, S. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat Protoc* **15,** 2247–2276 (2020).
34. Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A. & Lander, E. S. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176,** 928-943.e22 (2019).
35. Chu, B. K., Tse, M. J., Sato, R. R. & Read, E. L. Markov State Models of gene regulatory networks. *BMC Syst Biol* **11,** 14 (2017).
36. Shi, H., Song, L., Wu, Y., Shen, R., Zhang, C., Liao, X., Wang, Q. & Zhu, J. Edaravone Alleviates Traumatic Brain Injury by Inhibition of Ferroptosis via FSP1 Pathway. *Mol Neurobiol* (2024). doi:10.1007/s12035-024-04216-2
37. Wang, X., Shao, N., Zhang, X., Chen, H., Chang, Z., Xie, D. & Zhang, J. Ferulic Acid Activates SIRT1-Mediated Ferroptosis Signaling Pathway to Improve Cognition Dysfunction in Wilson's Disease. *Neuropsychiatric Disease and Treatment* **19,** 2681–2696 (2023).
38. Liu, G., Wang, T., Wang, T., Song, J. & Zhou, Z. Effects of apoptosis□related proteins caspase□3, Bax and Bcl□2 on cerebral ischemia rats. *Biomedical Reports* **1,** 861–867 (2013).
39. Zhou, Y., Liao, J., Mei, Z., Liu, X. & Ge, J. Insight into Crosstalk between Ferroptosis and Necroptosis: Novel Therapeutics in Ischemic Stroke. *Oxidative Medicine and Cellular Longevity* **2021,** 9991001 (2021).
40. Jindal, K., Adil, M. T., Yamaguchi, N., Yang, X., Wang, H. C., Kamimoto, K., Rivera-Gonzalez, G. C. & Morris, S. A. Single-cell lineage capture across genomic modalities with CellTag-multi reveals fate-specific gene regulatory changes. *Nat Biotechnol* **42,** 946–959 (2024).

41. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods* **14,** 414–416 (2017).
42. Jiang, J., Ye, X., Kong, Y., Guo, C., Zhang, M., Cao, F., Zhang, Y. & Pei, W. scLTdb: a comprehensive single-cell lineage tracing database. *Nucleic Acids Research* gkae913 (2024). doi:10.1093/nar/gkae913
43. Jiang, J., Li, J., Huang, S., Jiang, F., Liang, Y., Xu, X. & Wang, J. CACIMAR: cross-species analysis of cell identities, markers, regulations, and interactions using single-cell RNA sequencing data. *Briefings in Bioinformatics* **25,** bbae283 (2024).
44. Jiang, J., Lyu, P., Li, J., Huang, S., Tao, J., Blackshaw, S., Qian, J. & Wang, J. IReNA: Integrated regulatory network analysis of single-cell transcriptomes and chromatin accessibility profiles. *iScience* **25,** 105359 (2022).
45. Kong, Y., Jiang, J., Kong, W. & Qin, S. DRCTdb: disease-related cell type analysis to decode cell type effect and underlying regulatory mechanisms. *Commun Biol* **7,** 1205 (2024).
46. Jin, S., Plikus, M. V. & Nie, Q. CellChat for systematic analysis of cell–cell communication from single-cell transcriptomics. *Nat Protoc* (2024). doi:10.1038/s41596-024-01045-4
47. Longa, E. Z., Weinstein, P. R., Carlson, S. & Cummins, R. Reversible middle cerebral artery occlusion without craniectomy in rats. *Stroke* **20,** 84–91 (1989).