TOGGLE delineates fate and function within individual cell types via single cell transcriptomics

Junpeng Chen[1,2,16+*], Zhouweiyu Chen[3+], Tianda Sun[5+], Kaiqing Liu[1,2+], Eric Jiang[4], Yibing Nong[1], Tao Yuan[8,9], 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: Center for Cardiometabolic Science, Division of Environmental Medicine, Christina Lee Brown Envirome Insttitute, University of Louisville, Louisville, KY 40202, United States.
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: Tong Jiecheng Studio, Hunan University of Science and Technology, Xiangtan, 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: Tong Jiecheng Studio, Hunan University of Science and Technology, China
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: Department of Physiology, University of Louisville School of Medicine, Louisville, Kentucky,40204, USA
17 Address: Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing, China
18 Address: Kebei Hospital of Shaoxing

+ Junpeng Chen (drchenjunpeng@hotmail.com), 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 rankings are in no particular 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.

Biographical Note

Chen's guidance comes from Dr. Alex P. Carll. Another mentor is Dr. Lu Cai, one of the contributors to the standardization of autophagy detection methods, whose research on criteria for identifying diabetes induced apoptosis has been included in numerous clinical guidelines.

1.Introduction

RNA plays a central regulatory role throughout the life cycle of a cell—from formation and development to maturation, functional execution, aging, and pathological transformation. Dynamic changes in RNA expression precisely orchestrate cellular function at each stage. To capture these transitions, gene expression profiles across distinct cellular populations are typically compared. However, even within a single cell type, functional divergence or fate transitions may occur without significant transcriptomic differences, leading to a critical challenge: phenotypically similar cells may exist in vastly different functional states that are indistinguishable using standard expression-based methods. A salient example arises in ischemic stroke, where current single-cell RNA sequencing (scRNA-seq) approaches can distinguish broad neuronal subtypes (e.g., glutamatergic neurons) but fail to resolve neurons undergoing distinct stages of programmed cell death[1]. Similarly, neural stem cells exhibit marked DNA methylation shifts upon activation, yet their transcriptomic profiles remain nearly identical before and after transition, eluding traditional RNA-based classification[2].

We refer to this integrative methodology as Fate Tracing—a conceptual and computational framework for reconstructing the historical trajectories of functional transitions and programmed cellular

processes. It is necessary to use an unsupervised algorithm because the outcomes of costly cellular experiments are typically unknown in advance—otherwise, the algorithm would lose its purpose. Although numerous computational tools have been developed for scRNA-seq analysis, most are designed to model developmental trajectories or lineage differentiation, and offer limited resolution for identifying functional transitions within terminally differentiated cells, such as neurons. Current algorithms are primarily represented by WOT[3] and Cospar[4], Weighted Optimal Transport (WOT) relies on temporal information to infer cell trajectories but often confounds similar transitional paths. Cospar constructs state-transition models based on diffusion principles, yet lacks the capacity to resolve multi-stage classifications or fine-grained transitions within phenotypically stable populations. A shared limitation among these methods is their inability to distinguish between functionally distinct yet transcriptionally similar mature cell states.

To address these limitations, we developed TOGGLE, a computational framework tailored to delineate functional heterogeneity within mature cell populations. We conceptualize cellular function as a hierarchical system: core biological processes, such as metabolism and material transport, form the "trunk," while specialized functions emerge as branching "limbs." TOGGLE captures this hierarchy by applying a graph diffusion algorithm to construct multi-level representations of RNA regulatory networks. Coupled with a self-supervised learning scheme, the framework autonomously detects critical functional boundaries and enables fine-grained partitioning of cell populations. The resulting diffusion-derived modules reveal transcriptional specialization that traditional clustering methods fail to capture. TOGGLE thereby establishes a high-resolution functional hierarchy that facilitates discrimination not only between broad cell classes but also among subtle functional subtypes. It further disentangles overlapping biological programs that obscure classification boundaries, enabling precise identification of key transitional states along cell fate trajectories. To operationalize this approach, we constructed an RNA-based functional graph model for visualizing and quantifying transcriptional function.

2.Methods

2.1 Dataset & How to Verification

2.1.1 TEST 1

To evaluate the accuracy of our method, we first applied it to several publicly available scRNA-seq datasets with clearly annotated ground-truth outcomes.

(1) GSE140802[5]: Mouse bone marrow dataset, a subset of progenitor cells differentiates into monocytes—immune cells specialized in pathogen engulfment—while others give rise to neutrophils,

another distinct immune lineage. These are referred to as monocyte progenitors and neutrophil progenitors, respectively.

(2) GSE99915[6]: Reprogram mouse dermal cells into endoderm-like progenitors. A fraction of cells successfully acquired the target identity (reprogrammed progenitor cells), whereas others failed to transition (failed progenitor cells).

Since the final cellular fates are explicitly annotated in both datasets, they provide robust benchmarks for evaluating classification and fate-prediction performance.

2.1.2 TEST 2

We then applied this method to a scRNA-seq dataset without cell state annotation.

GSE232429[1]: A dataset of neurons after ischemic stroke. It was used to identify cells from early, intermediate, and late stages of programmed cell death.

To evaluate the reliability of the prediction results, We observed significant alterations in genes involved in apoptosis and ferroptosis (two forms of programmed cell death) in the relevant cell populations, which can accelerate programmed cell death. These transcriptomic signatures were further corroborated by animal experiments, in which consistent protein expression changes were detected in the ischemic brain tissue of rats.

2.1.3 TEST 3

We further assessed its functional biology discovery capabilities.

GSE209656[2,7]: Identified quiescent neural stem cells (qNSCs) embedded within a broader population of glial cells—a particularly challenging task due to the near-identical transcriptomic profiles shared by qNSCs and astrocytes.

By leveraging the Graph Diffusion Functional Map, we clustered RNAs based on functional similarity and uncovered epigenetic signatures associated with energy metabolism regulation.

2.1.4 TEST 4

These results are detailed in the Supplementary Appendix 1(**RNA Pathway Alterations in Air Pollution**)&2(**Cellular functional mapping**).

(1) GSE183614[8]: it was used to detect RNA pathway perturbations.

(2) GSE253768: it was used to analyze intercellular communication.

2.2 Method for distinguishing cellular functions and functional phases

The model consists of three modules. Section 2.2.1 describes the graph scattering module, which converts the sequencing dataset into an expression matrix and computes pairwise similarities between cells to serve as the foundation for downstream classification analysis. Section 2.2.2 introduces a

self-supervised module (a kind of unsupervised learning) integrated with reinforcement learning principles. This enables the model to automatically classify cells without predefined labels and iteratively refine itself, ultimately generating clusters of functionally similar cells. These clusters represent distinct stages within functional states. To further resolve these stages, a masking optimization module is introduced in Section 2.2.3 to identify and consolidate cells sharing similar functional characteristics. The overall workflow of the algorithm is illustrated in Figure 1.

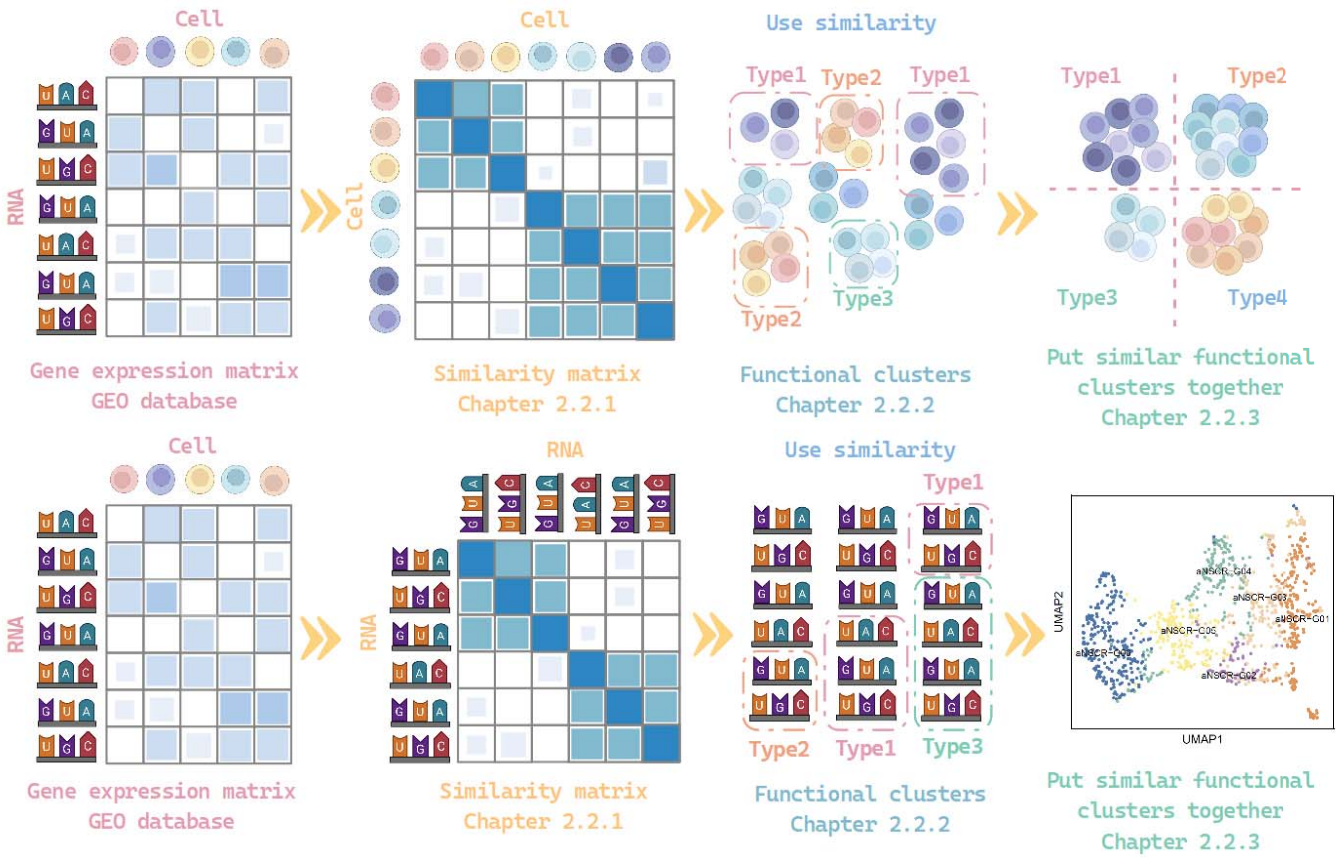


Figure 1. **Top**: Derivation of Cell Functions from a Gene Matrix. **Bottom**: Derivation of RNA functions from a Gene Matrix. A **gene expression matrix** is used to retrieve RNA transcript counts for individual cells from the GEO database. A **similarity matrix** calculates pairwise similarities between cells to enable classification. **Functional clusters** identify initial groups of cells with similar biological functions. And lastly, similar functional clusters are merged together.

2.2.1 Step 1: Generate intercellular similarity using a graph diffusion algorithm from deep learning

Traditional cell similarity measures rely on direct gene expression comparisons, which can miss subtle functional relationships between cells. To address this limitation, TOGGLE employs a graph diffusion algorithm that captures higher-order relationships by considering how gene expression patterns propagate through cellular neighborhoods, revealing functional similarities that may not be apparent

from direct expression comparisons. We utilize single-cell RNA sequencing (scRNA-seq) expression data as input, with example datasets sourced from the Gene Expression Omnibus database (accession No. GSE140802, GSE99915)[1,5,9–12]. The graph diffusion matrix exhibits asymmetry, where the upper and lower halves obtained by diagonal splitting contain different functional information. This approach utilizes intercellular graph distances for similarity measurement, incorporating spatial distribution information where 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 2).

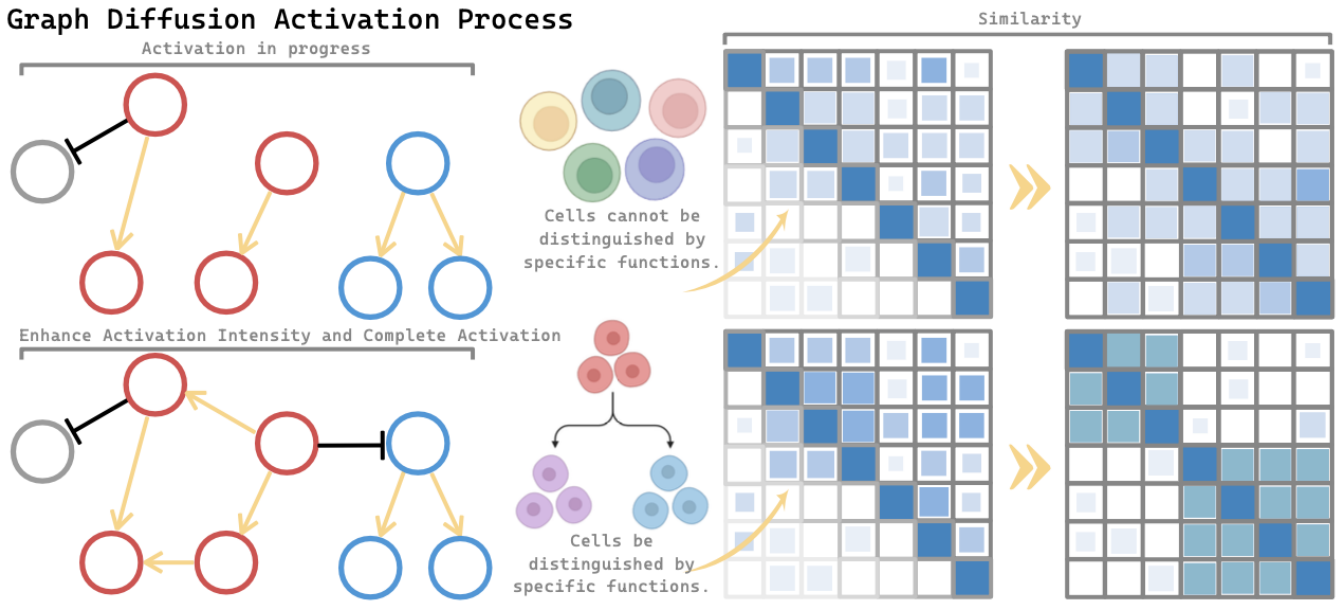


Figure 2. Top panel: In conventional models (left), the initial model activation is limited. Due to the masking effect of similar gene expression profiles (middle), functional differences between cells remain indistinct, resulting in poor separation in the data space (right). **Bottom panel:** By applying TOGGLE's novel framework, deeper model activation is achieved (left). Cells are distinguished based on both functional identity and progression stage (middle), leading to more pronounced separation in the data space (right). The data separation is visibly more distinct.

To obtain an RNA-seq expression matrix of size , transcriptional results for RNA molecules from cells were retrieved from the NCBI database, resulting in a non-symmetric graph diffusion matrix of size . This matrix provides information about cell similarity. Extracting the row and column of the matrix yields , representing the one-way graph distance from the cell to the cell. Due to the asymmetry of the scatterplot distances, represents the one-way graph distance from the cell to the cell, typically not equal to .

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 , clusters 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 $\overline{D_i}$ for the similarity matrix M, allowing the data to be measured around the assumed virtual origin point.

$$\overline{D_i} = D_i - \frac{1}{n}\sum_{j=1}^{n} D_{ij}$$

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 $\overline{D}$ to enhance the dissimilarity in the similarity matrix, thereby mapping the original data to a latent space with greater differences.

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

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

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

Where $diag(cov_D)$ 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 M, represented in the form of the correlation assumption ∂. ∂_n represents the $n-th$ row vector in the matrix, with correlation coefficients for sample cells $-1 < \partial_n < 1$. Based on the numerical characteristics of the singular matrix, during the mapping process from the original object M to the mapped object pM ($M \rightarrow pM$), 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 D‘. In simple terms, using $pM = \gamma\ (M)$ amplifies the differences within the similarity matrix M, making it easier to classify elements within the matrix.

2.2.2 Step 2: Identify functionally similar cell clusters using a self-supervised learning algorithm

We first processed the raw data into gene expression matrices, capturing the expression levels of each gene across individual cells. Following the procedure described in Section 2.2.1, we computed pairwise similarities between cells to provide a foundation for downstream classification. Subsequently, using the strategy outlined in Section 2.2.2, we identified clusters of functionally similar cells. This step integrated principles from self-supervised and reinforcement learning to construct an unsupervised algorithm that requires no predefined labels. The model autonomously generates pseudo-labels and iteratively trains itself to detect functional states. Finally, as described in Section 2.2.3, we further grouped cells that shared similar functional roles but existed at different stages of functional progression, enabling refined biological difference analysis (Figure 3).

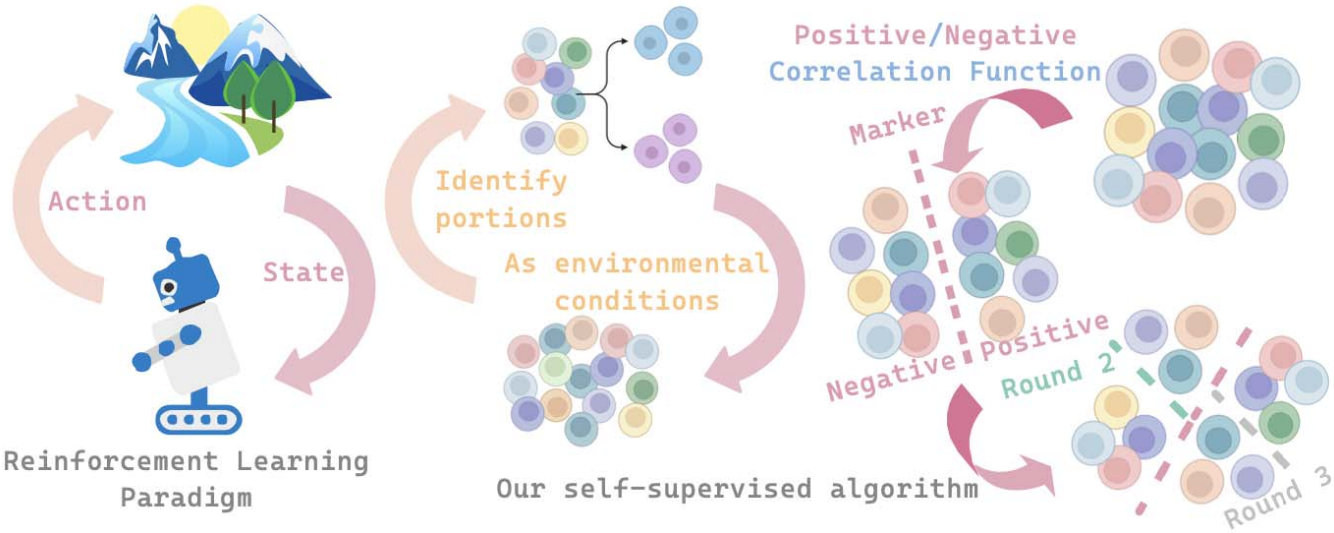


Figure 3. Right: The algorithm first generates a mathematical condition used to evaluate whether each cell is positively or negatively correlated. The boundary separating these two categories is defined as a "Marker." Each pair of adjacent Markers delineates a cluster, comprising cells that exhibit similar functional characteristics. This process requires no manual labeling; the algorithm autonomously generates and refines Markers through iterative feedback, enabling unsupervised functional clustering.
Left: Conceptual illustration of reinforcement learning principles. **Center:** The algorithm treats the task of identifying spatial regions of cell distribution as an output action (Action), which is then fed back as the subsequent state or environmental condition (State/Environment).

When directly using correlation coefficients for classification, there is a presence of numerous boundary samples between different clusters, which makes the boundaries less distinct or prone to being misclassified as independent clusters. Assuming that is located at the clustering center of cluster , when the convergence of cluster within its hypothetical space occurs, there exists a loss function, denoted as :

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 D', the knowledge reservoir used consists of D^+ and D^-, 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.

2.2.3 Step 3: Group functionally identical cell clusters together using the Masking principle.

The correlation-based clustering from Step 2 generates boundary markers that separate different functional regions, making it straightforward to determine each cell's relative position with respect to these markers. By selecting cells located between adjacent markers, we can identify functionally similar

cell groups. However, cellular functional relationships are often nonlinear, meaning that different cell types with distinct biological functions may still exhibit similar correlation patterns in the similarity space. This poses a fundamental limitation: correlation-based grouping alone is insufficient to determine final cell type classifications, as it may merge functionally distinct cell populations that happen to share similar correlation signatures. To address this challenge, we developed a Deep Genetic Network that adapts the masking framework from BERT to an unsupervised setting. This approach captures global similarity relationships between individual cell clusters and all other clusters, enabling more precise functional classification by considering contextual information beyond local correlations. The final functional assignment of each cluster is determined through this integrated framework (Figure 4).

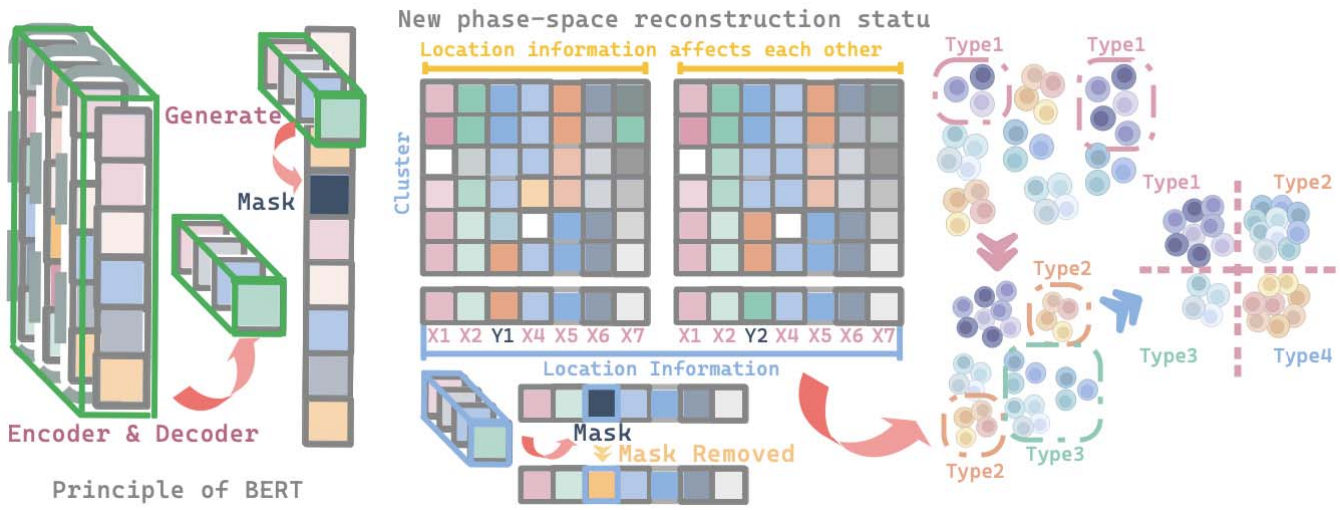


Figure 4. Right: (1) In Section 2.2.2, the algorithm brings together cells with the same functional type (Type 1), even if they are in different conditions. (2) In Section 2.2.3, it further merges all cells with Functional Type 1 into a unified group. (3) Eventually, all cells with the same functional type—regardless of condition—are grouped together.
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.

After completing the calculations in the previous section, the sorted matrix is obtained. The internal boundary of is , where is the boundary generated during the -th iteration. is the predictive information obtained through unsupervised learning. After using 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, setting $Y = rM$ allows the conclusion $E(Y_i, \varsigma) \Leftrightarrow E(Y_i)$ to be applied.

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[13], 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.

2.3 Another function: RNA Functional Differentiation Map (Graph Diffusion Functional Map)

We used UMAP to visualize the functional grouping of RNA. While UMAP effectively captures overall expression trends, its dimensionality reduction process inevitably results in information loss, leading to the mixing of RNA species with distinct functional roles. This mixing effect hinders accurate discrimination of functional states and limits the depth of target identification and functional interpretation. To address this limitation, we propose a novel approach: the Graph Diffusion Functional Map. As described in Section 2.2, we replace the original cell-based graph diffusion matrix M_{ij} with a matrix derived from RNA–RNA correlations and apply the full computational framework detailed in the Methods section to construct the functional map. This strategy significantly reduces functional mixing while preserving the intrinsic correlation structure among RNAs, enabling clearer delineation of RNA functional groupings. Unlike conventional dimensionality reduction techniques, this graph diffusion-based method retains high-dimensional information and improves sensitivity to subtle functional differences. Our results demonstrate that the Graph Diffusion Functional Map not only recapitulates known functionally enriched clusters but also reveals previously undetected, highly expressed RNA groups overlooked by UMAP. These newly identified targets provide novel insights into single-cell transcriptomic data and expand opportunities for understanding disease mechanisms and identifying therapeutic targets.

2.4 scRNA-seq Data Processing and Experimental Methods

Details of scRNA-seq data processing and experimental methods are shown in the Supplementary Appendix 6 – Animal Processing.

3.Result

3.1 Test 1: Performance Testing Using Developmental Cells

TOGGLE is an unsupervised algorithm and does not require any prior training. To evaluate its

performance, we compared TOGGLE with four widely used computational methods: Cospar[4], Super OT[14], WOT[3], Deep-Lineage[15,16], and GAN-Based OT[13].

Mouse Bone Marrow Hematopoietic Progenitor Cells (Figure 5A-D):

- When evaluating only the progenitor cells, TOGGLE achieved accuracy of 79%. Notably, Cospar achieved an accuracy of 65%, Super OT slightly higher at 67%, WOT around 65%, and GAN-Based OT at 62%, Deep-Lineage (Supervised) at 52%.
- When evaluating progenitor and descendant cells, TOGGLE achieved an accuracy of 82%. In contrast, WOT's accuracy was below 50%. Because no cells' preset information is provided, other methods were not applicable for this setting.

Lineage-Reprogrammed Cells (Figure 5E-H):

- When evaluating only the progenitor cells, TOGGLE achieved 91.2%. As an unsupervised algorithm, TOGGLE was benchmarked against the two most common computational methods, with Cospar achieving 67% accuracy and WOT only 31%, Deep-Lineage at 51%, demonstrating TOGGLE's significant performance advantage. Accuracy of other methods was below 50%.
- When evaluating progenitor and descendant cells, TOGGLE achieved an accuracy of 84%, while WOT remained below 50%. Because no cells' preset information is provided, other methods were not applicable for this setting.

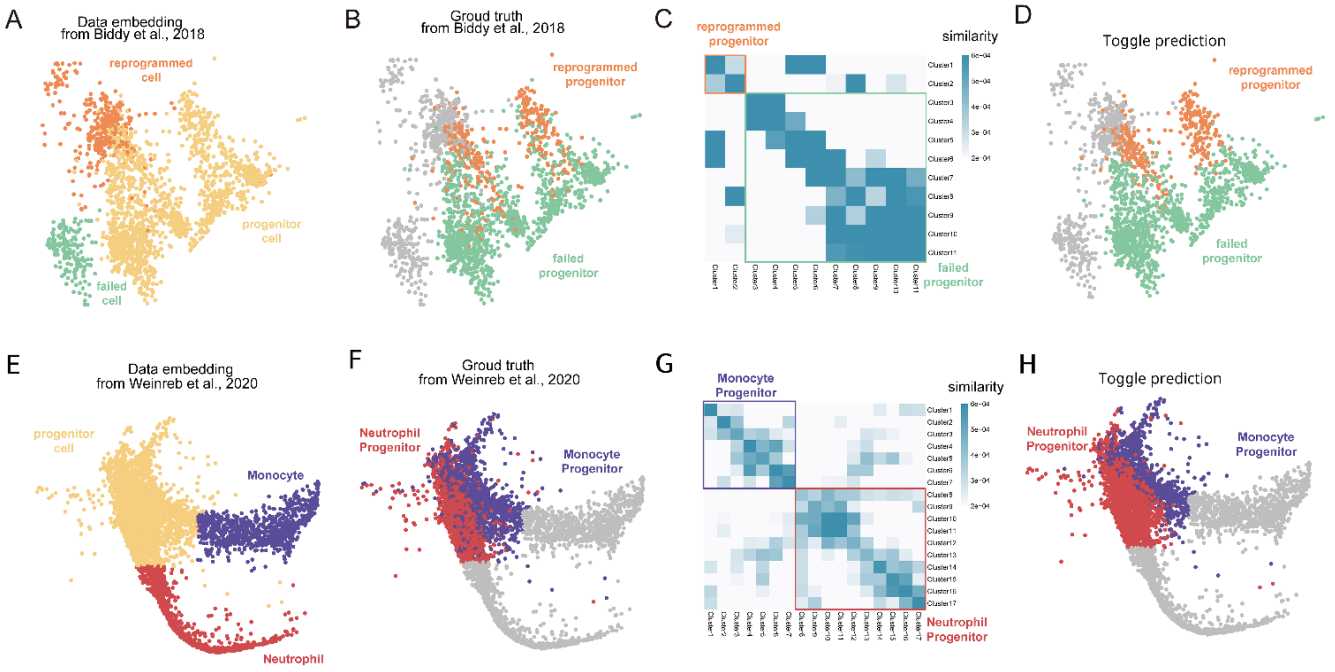


Figure 5. Tracing the lineage of fibroblast reprogramming and hematopoietic cell development. A/E: UMAP visualization of the original dataset. The Progenitor Cell population represents the targets to be predicted. **B/F:** Real label of lineage outcomes showing the eventual fate of

Progenitor Cells. **C/G:** Algorithm-predicted results with output cluster boundaries. **D/H:** Final fate predictions based on the inferred trajectories.

3.2 Test 2: Distinguish the stages of ferroptosis and perform biological validation.

Detailed protocols for all animal experiments are provided and can be found in the **Supplementary Appendix 3**. We utilized the GSE232429 dataset to identify and characterize neuronal cell death associated with ischemic stroke, including both middle cerebral artery occlusion (MCAO) and healthy control (Sham) samples. Throughout the study, MCAO and Sham samples were combined and submitted to the algorithm for integrated clustering, followed by separate visualization based on their respective labels.

- Initially, 5 groups were automatically identified (Figure 6A). Among them, R1-5/4/3 exhibited no significant changes and were considered not directly associated with cell death. Through pairwise comparison of differentially expressed genes related to programmed cell death, R1-1/5 were determined to represent healthy states, whereas R1-2 corresponded to a death-related state, and R1-3/4 were classified as intermediate states (Figure 6B). Subsequently, R1-1 was selected as the reference group due to its direct continuity with R1-5. R1-2 and R1-5 were included in the subsequent analysis.
- In the 2 round of classification, cells were further subdivided into 9 subpopulations (Figure 6C). Among them, R2-8/9 appeared in both Sham and MCAO groups, so they were identified as normal cells. In contrast, R2-2/3 were markedly enriched in the MCAO group and were closely associated with programmed cell death[17,18].
- Using R2-9 and R2-8 as the reference group for differential gene analysis, we found that R2-3 was most enriched for signaling pathways and genes associated with programmed cell death (Figure 6D). Other groups represented intermediate states with relative low-level expression of programmed cell death-related genes (Figure 6D) [17,18].
- The populations R2-2/3 were used for 3 round of clustering (Figure 6E). Based on the criteria of Log2FC $\geq$ 0.25 and FDR < 0.05, genes that were specifically and highly expressed in each subgroup were identified (Figure 6F). Among the differentially expressed genes in Groups R3-4/5, a large number of ferroptosis driver genes were found to be highly expressed, with over 70% of cells exhibiting ferroptosis markers (Figure 6G). Enrichment analysis further revealed that the differentially expressed genes in this cell group were related to ferroptosis, hence, it was speculated that the main form of programmed cell death in R3-4/5 was ferroptosis[17,18].

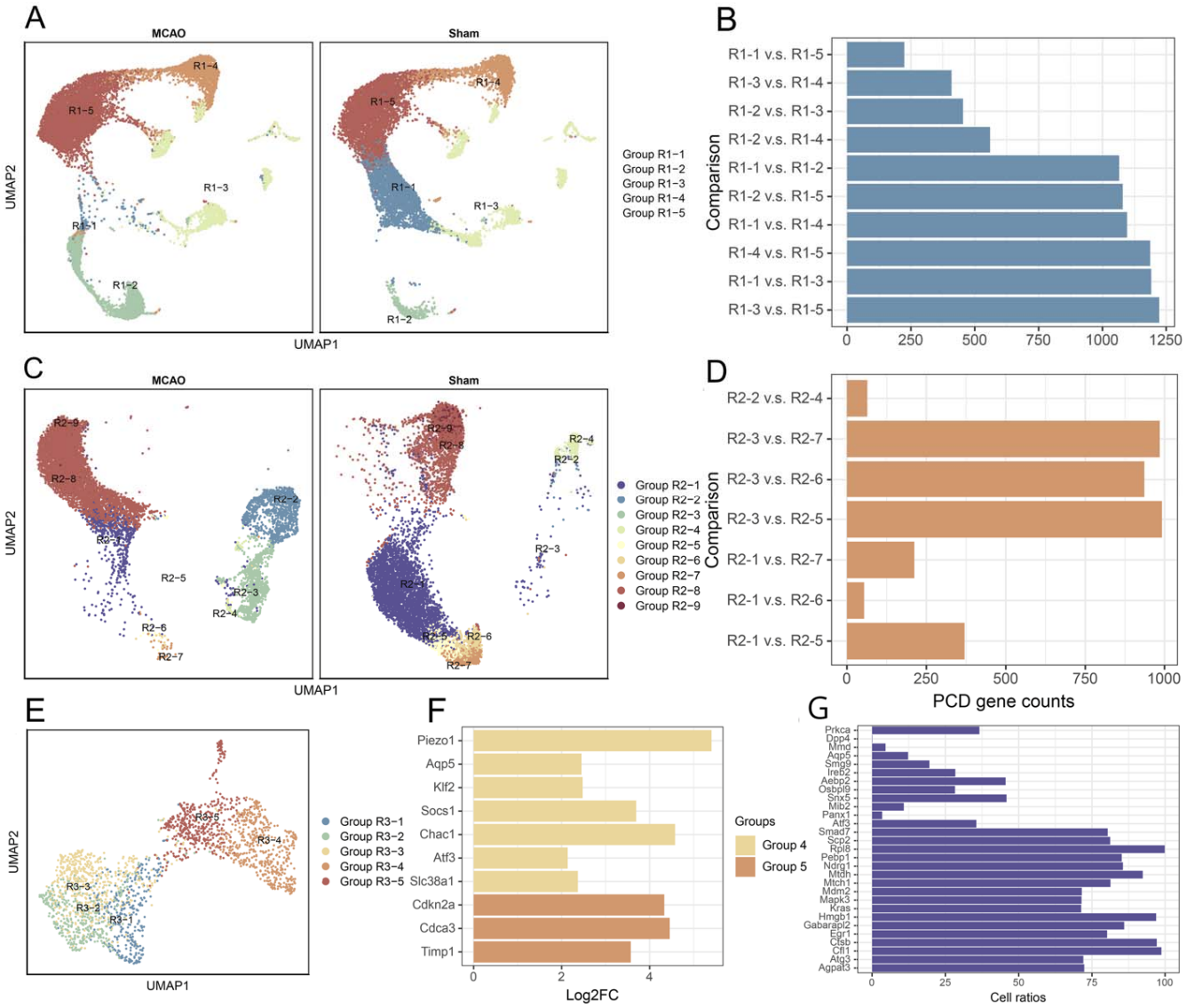


Figure 6. In this study, middle cerebral artery occlusion (MCAO) and healthy control (Sham) samples were combined in clustering analysis, then independent plots were generated based on MCAO and Sham labels. **Round 1 (A):** R1-5 is the corresponding healthy cell control group, while R1-1 and R1-2 are related to cell death. **Evidence of differential gene count in round 1 (B):** Comparison of the number of differentially expressed programmed cell death genes between cell groups of all cohorts. The fewer the number of differentially expressed genes, the more similar the states of the two cell groups are. It can be seen that R1-2 is significantly dissimilar to other groups. **Round 2 (C):** R2-3 is identified as a cell population approaching death. R2-2 and R2-1 are in intermediate state containing a certain level of death-related gene expression. **Evidence of differential gene count in round 2 (D):** It can be seen that R2-3 is significantly dissimilar to other groups. **Round 3 (E):** R2-2 and R2-3 were selected and divided into 5 groups. **Log2FC values (F):** The larger the Log2FC value, the higher the gene expression. **Evidence of differential gene count in round 3 (G):** Enrichment analysis identified R3-4 and R3-5 as ferroptotic cell populations, with ferroptosis marker genes expressed in more than 70% of cells in these groups.

3.2.1 Verification of apoptosis- and ferroptosis-related genes

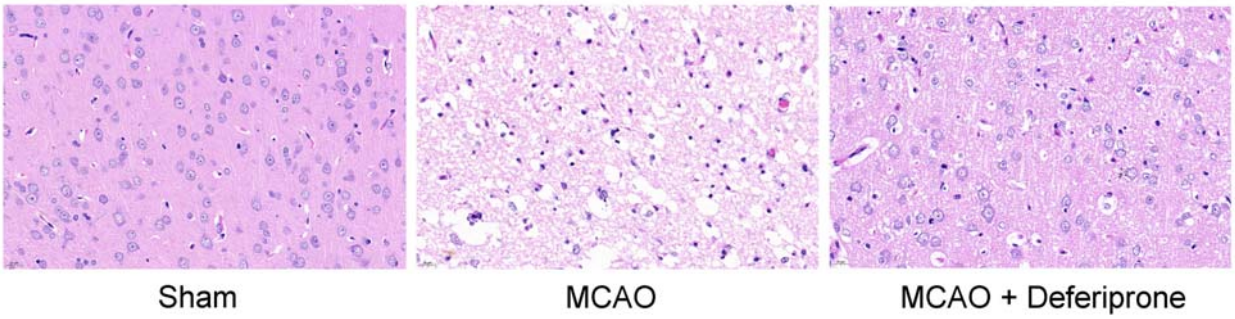
The complete animal experimental protocols are provided in Supplementary Appendix 6 – Animal Processing. Our previous research demonstrated increased iron levels and neuronal iron accumulation in the infarcted brain tissue of the MCAO model. Deferiprone treatment in MCAO model effectively reduces iron content and diminishes iron accumulation[19]. Based on these findings, we used the MCAO model ± deferiprone to validate the predicted results of neuronal ferroptosis by TOGGLE.

3.2.2 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 7).

3.2.3 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 7).



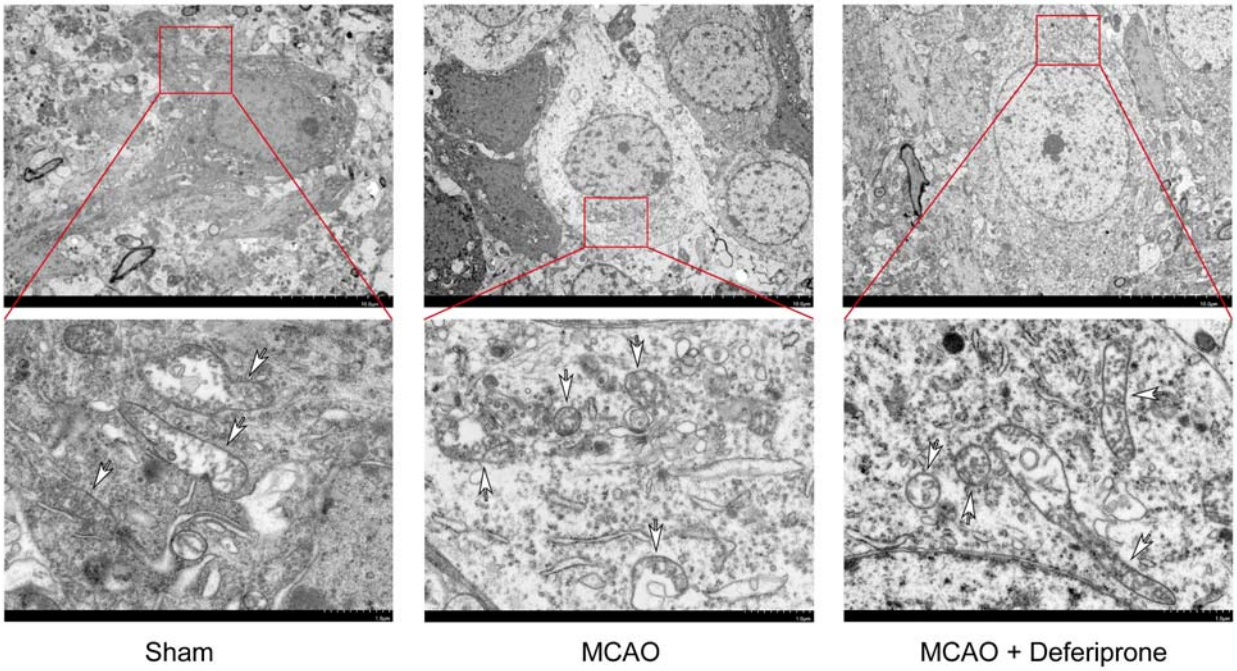


Figure 7. Pathological changes of brain tissue (Top). HE staining, 40× magnification. **Ultrastructural changes of brain tissue (Bottom).** Transmission electron microscope. The upper images are at 1500× magnification, and the lower images are at 6000× magnification. Mitochondria are marked by white arrows. **Summary:** Through these results, we observed characteristic cellular ultrastructural changes of ferroptosis.

3.2.4 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 8).

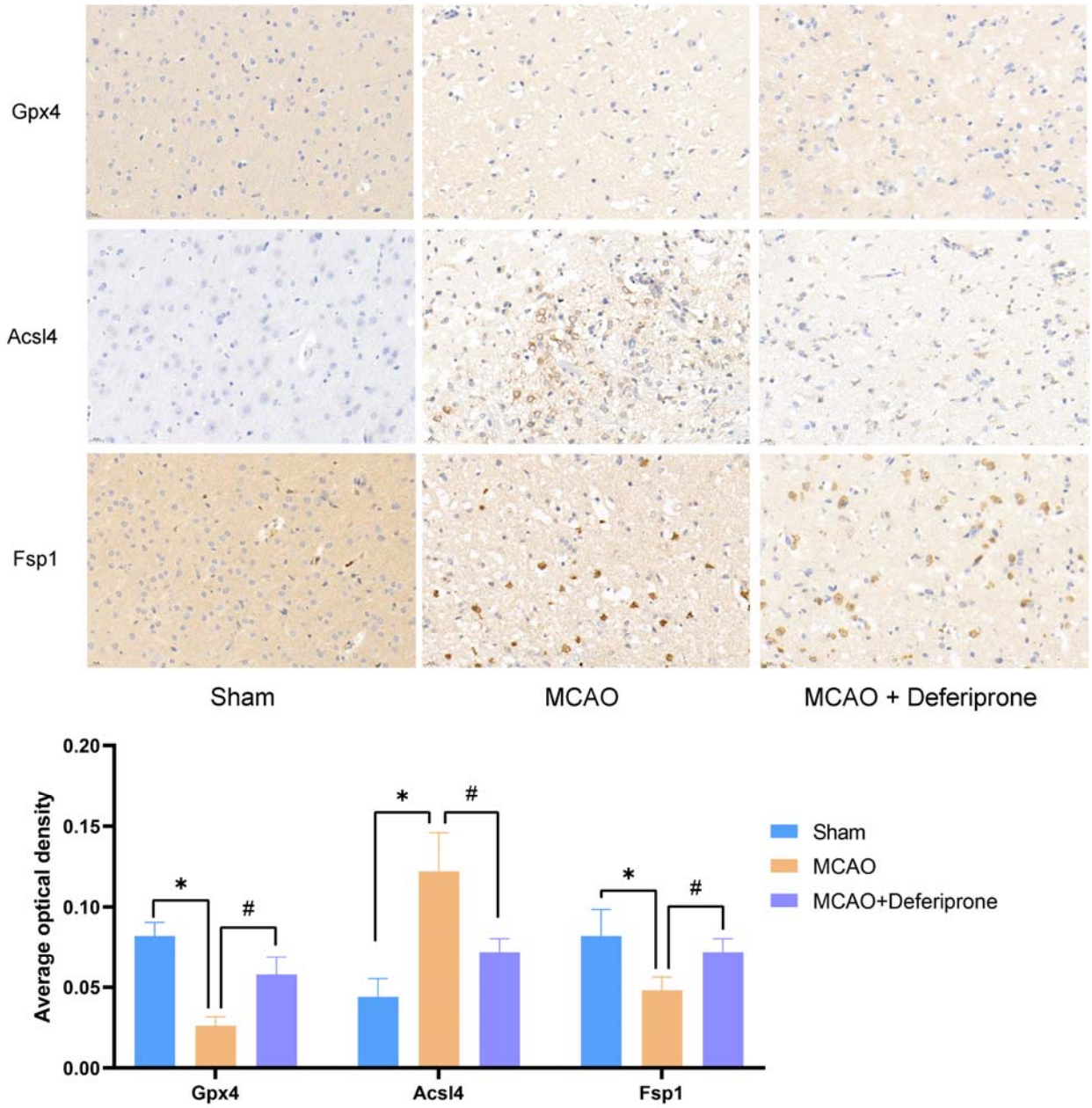


Figure 8 Expression levels of Acsl4 and Gpx4 proteins. Immunohistochemistry, 40× magnification. *compared with the sham group, P<0.05; #compared with MCAO group, #P<0.05. n=3 per group

3.2.5 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 9).

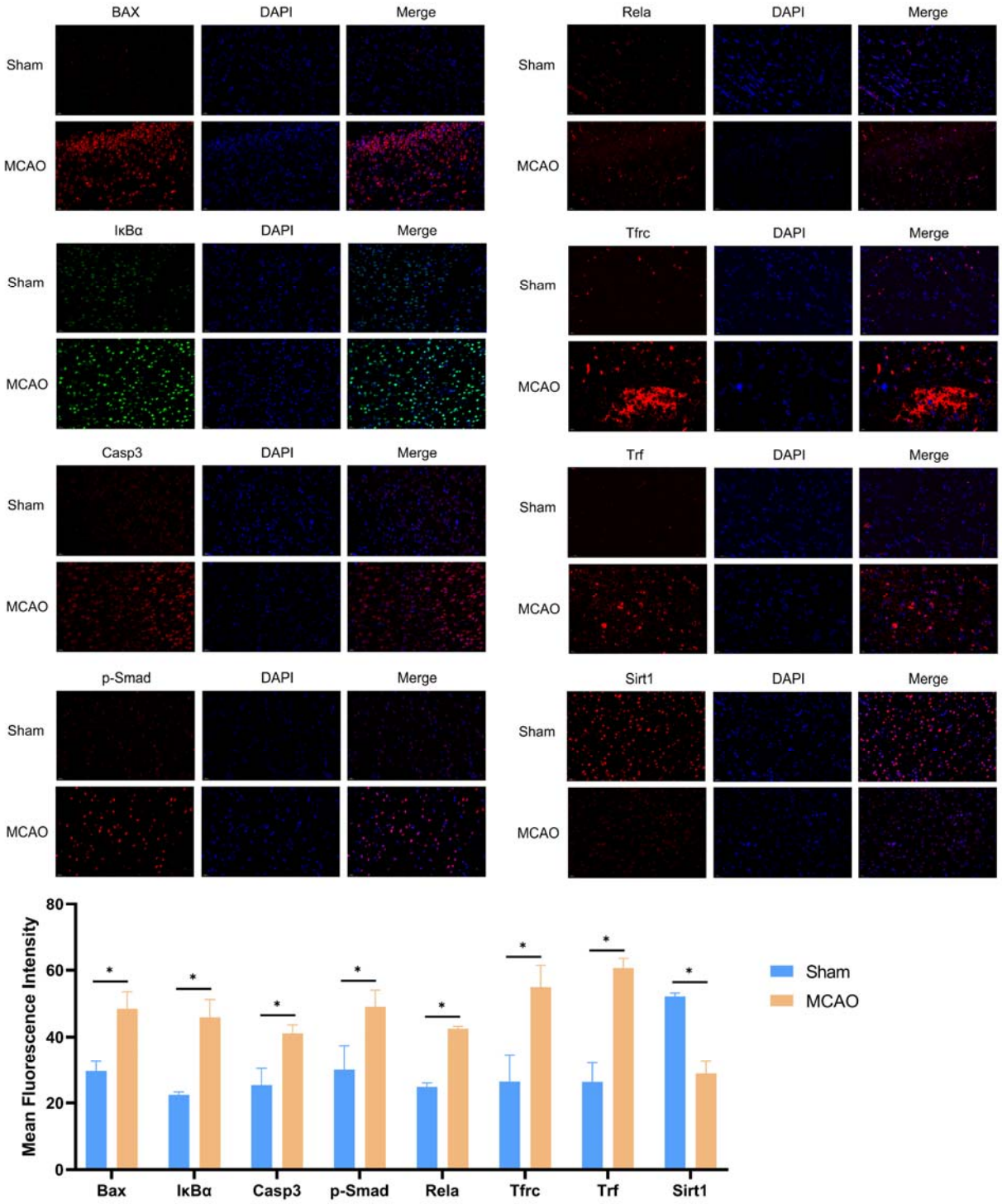


Figure 9 Expression levels of Bax, IkBa, p-Smad, p65, Tfrc, and Trf proteins. Immunofluorescence, 40× magnification. Compared with the sham group, *P<0.05. n=3 per group.

3.3 Test 3: Identifying Epigenetic Metabolic Regulation in NSCs Using Novel Functional Mapping

3.3.1 Establish New RNA Functional Map (Graph Diffusion Functional Map)

We used Groups R3-4 and R3-5 from Section 3.2 to evaluate the effectiveness of RNA functional clustering. In the GSE209656 dataset (Figure 10), the top 1,000 most variable genes were selected from active neural stem cells (aNSCs) and quiescent neural stem cells (qNSCs). The results obtained using the method in Section 2.3 are shown in Figure 11a-f and Table 1.

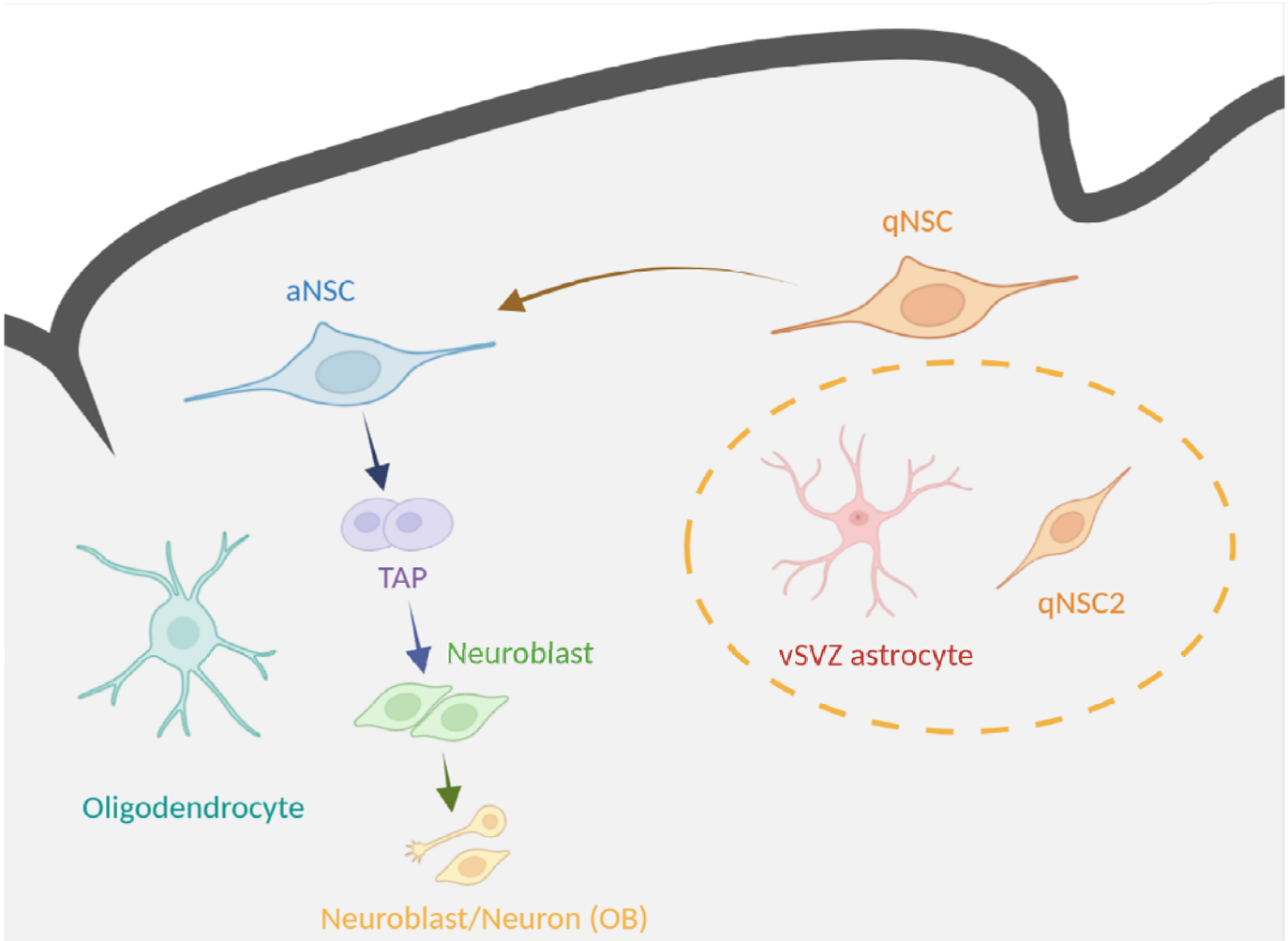


Figure 10 Brief background: Schematic diagram of the vSVZ adult NSC lineage. Referenced from Kremer et al. 2024[2]. qNSC: quiescent neural stem cell; qNSC2: Primed quiescent neural stem cells; TAP: transient amplifying progenitors; OB: olfactory bulb. qNSCs are classified into — qNSC2 and vSVZ. qNSCs can transition into aNSCs, and aNSCs can further differentiate into other types based on functional states. The complexity of these relationships renders most existing algorithms ineffective. Created by Biorender.

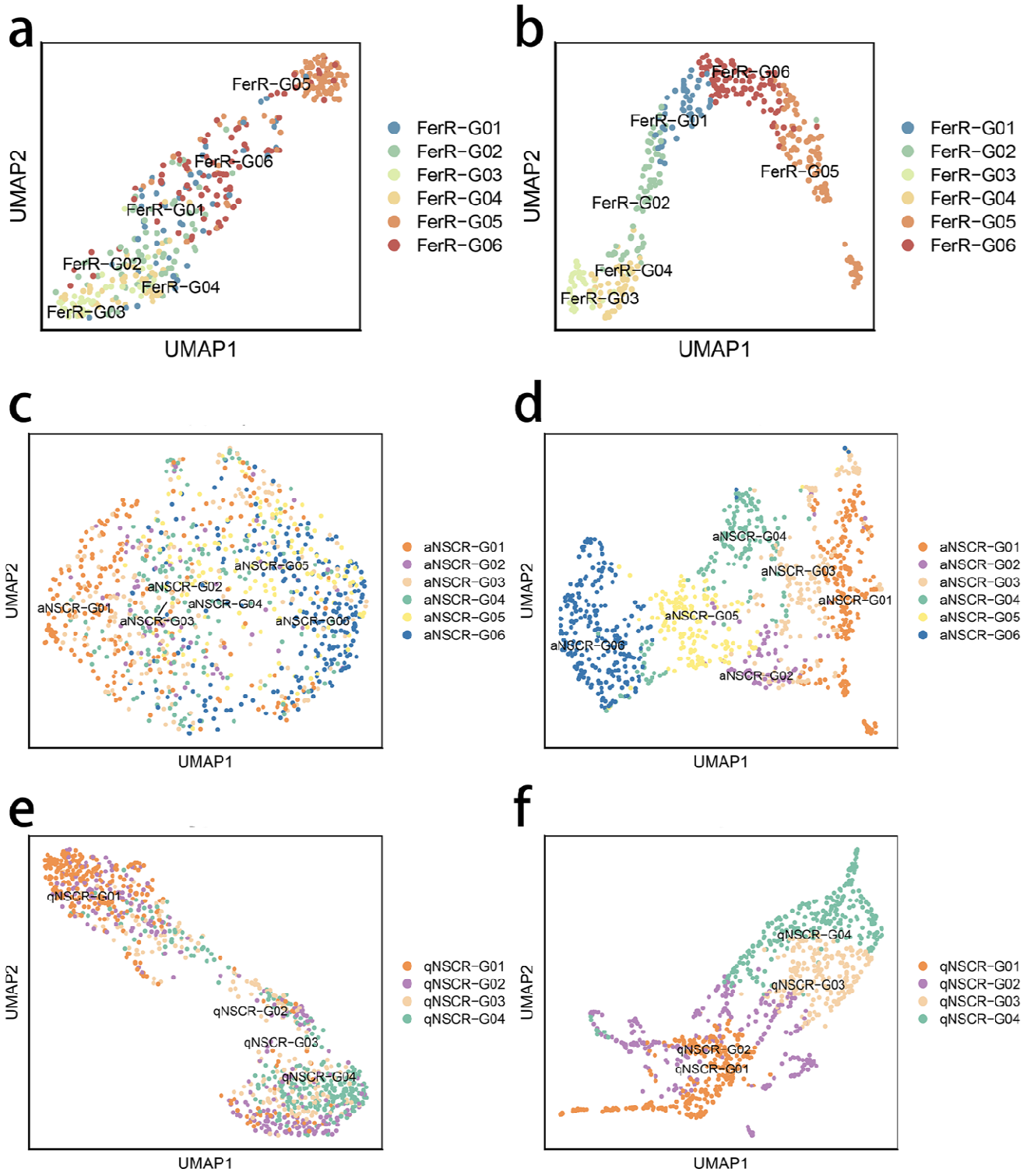


Figure 11 RNA functional profiles from different data. a: RNA map of ferroptosis-related genes in Group R3-4 and R3-5 cells made by using single-cell expression matrix; b: RNA map made using the new method for the data in a; c: RNA map of 1000 highly variable genes in aNSC cells made by using single-cell 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.

Table 1: RNA Functions of figure 11a-f

Group Number	Functions
FerR-G01	cellular response to starvation (GO:0009267) cellular response to hypoxia (GO:0071456) positive regulation of blood vessel endothelial cell migration (GO:0043536), positive regulation of cell migration (GO:0030335)
FerR-G02	negative regulation of apoptotic processes (GO:0043066) cellular response to oxidative stress (GO:0034599)
FerR-G03	associated with autophagy (GO:0006914) autophagosome maturation, and negative regulation of transcription by RNA polymerase II (GO:0000122)
FerR-G04	cell redox homeostasis (GO:0045454) regulation of the cell cycle (GO:0051726) and the endoplasmic reticulum unfolded protein response (GO:0030968)
FerR-G05	superoxide metabolic processes (GO:0006801) inflammatory responses (GO:0006954) cellular response to hydrogen peroxide (GO:0070301)
FerR-G06	reactive oxygen species (GO:0034614) SMAD protein signal transduction (GO:0060395)
aNSCR-G01	cell differentiation (GO:0030154)
aNSCR-G02	transcription by RNA polymerase II (GO:0006357)
aNSCR-G03	chromosome segregation (GO:0007059) proteolysis (GO:0006508)
aNSCR-G04	DNA repair (GO:0006281) cell division (GO:0051301)
aNSCR-G05	cell migration (GO:0016477) regulation of gene expression (GO:0010468) positive regulation of cell growth (GO:0030307)
aNSCR-G06	mRNA processing (GO:0006397) RNA splicing (GO:0008380) chromatin remodeling (GO:0006338)
qNSCR-G01	centriole replication (GO:0007099) regulation of the mitotic cell cycle (GO:0007346)
qNSCR-G02	cell adhesion (GO:0007155) apoptotic process (GO:0006915)
qNSCR-G03	RNA polymerase II (GO:0006357) endoplasmic reticulum to Golgi vesicle-mediated transport (GO:0006888)
qNSCR-G04	proton motive force-driven mitochondrial ATP synthesis (GO:0042776) proton transmembrane transport (GO:1902600)

The interpretation of the results is provided in: Supplementary Appendix 4/5.

We performed functional enrichment analysis on RNA clusters derived from ferroptosis-related cell populations (FerR-G01 to G06) and neural stem cells (aNSCs and qNSCs), focusing on highly variable genes. Taking ferroptosis as an example (Figure 12), FerR-G01 was enriched in hypoxia response and pro-angiogenic functions. These enriched pathways suggest that this group may help alleviate hypoxia in

ischemic brain regions by promoting neovascularization. Key genes such as Nfs1, Osbpl9, and Pgd may regulate iron metabolism, lipid signaling, and antioxidant capacity. Genes not enriched in specific pathways may still be functionally associated with their RNA group, potentially revealing novel mechanisms of ferroptosis. For example, in FerR-G01, Nfs1 plays a critical role in the biosynthesis of iron–sulfur (Fe–S) clusters, which are essential for mitochondrial respiratory chain complexes. Hypoxia caused by cerebral ischemia can impair mitochondrial function, reduce ATP production, and lead to cellular injury. Nfs1 may regulate ferroptosis by maintaining Fe–S cluster homeostasis, thereby influencing iron metabolism and antioxidant defense. Osbpl9 is a key regulator of cholesterol and phospholipid exchange, involved in intracellular lipid transport and signal transduction. It may promote endothelial cell migration and angiogenesis in ischemic regions by modulating the lipid microenvironment and associated signaling pathways. Pgd participates in the pentose phosphate pathway (PPP) and primarily functions by generating NADPH to maintain cellular antioxidant capacity. Under hypoxic conditions, elevated reactive oxygen species (ROS) levels induce oxidative stress. Pgd may counteract ferroptosis by boosting antioxidant defenses, protecting cells from oxidative damage.

The complete subgrouping results and target discovery details are provided in the Supplementary Appendix 3 – Full details of ferroptosis analysis.

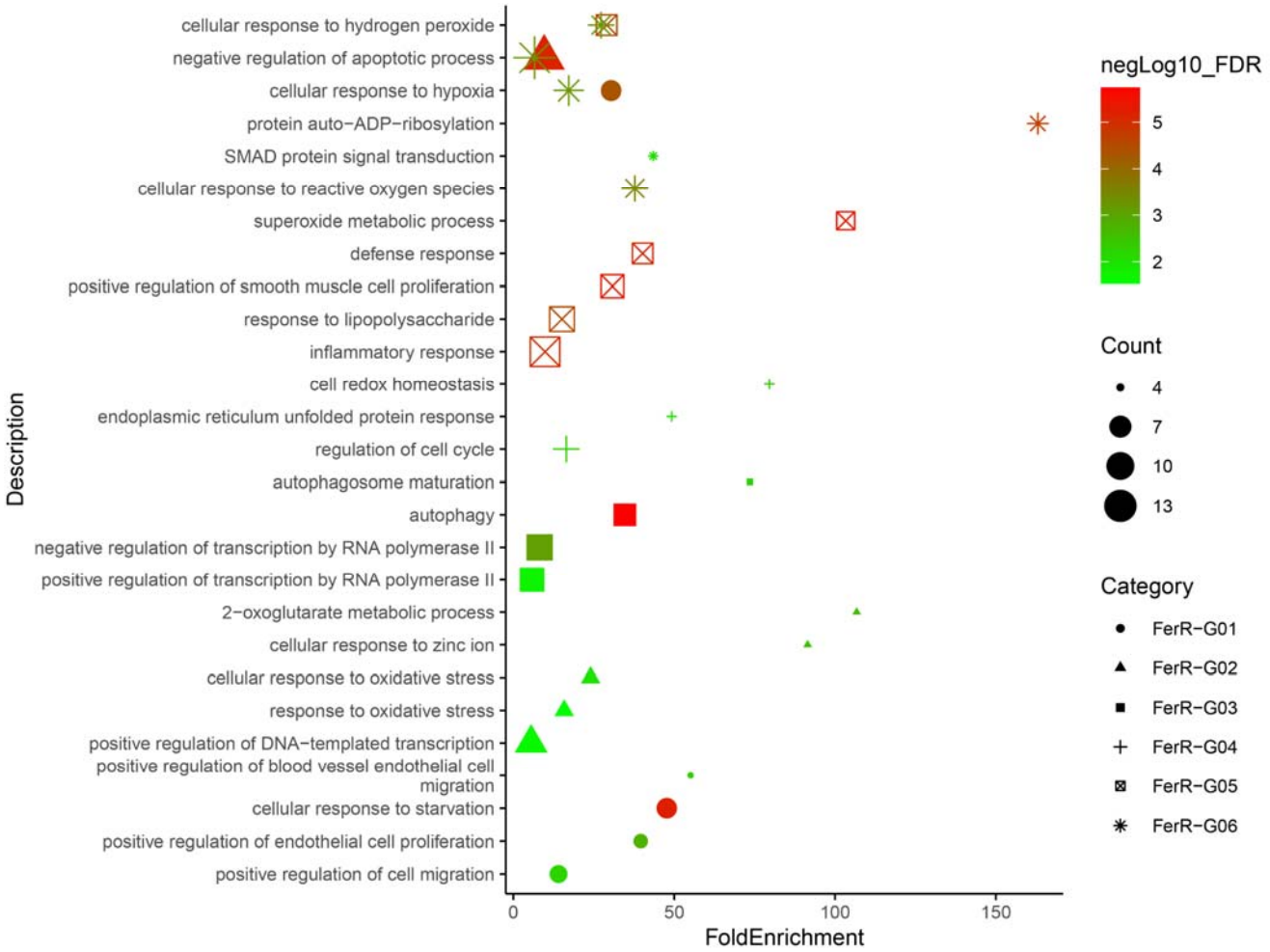


Figure 12 Bubble chart: The function of each RNA group in Figure 11(B), as determined by enrichment analysis. Genes not enriched in specific signaling pathways may still be functionally associated with their corresponding RNA groups, potentially revealing novel mechanisms involved in ferroptosis.

3.3.2 Identifying Epigenetic Metabolic Regulation in NSCs Using Novel Functional Mapping

Neural stemness refers to the ability of neural cells to self-renew and differentiate into other cell types. This state is not static but continuously evolving, which blurs the functional boundaries between different neuronal states and makes accurate classification of these states extremely challenging.

Kremer et al[2]. demonstrated that despite their distinct functional roles, astrocytes and quiescent neural stem cells (NSCs) exhibit highly similar transcriptomic profiles[2]. However, studies have shown that under specific conditions—such as injury or overexpression of key factors—astrocytes can reactivate their latent neurogenic potential. This discovery opens new avenues for regenerative medicine: by regulating cellular memory and energy metabolism, it may be possible to induce astrocytes to acquire neurogenic capacity, enabling regenerated tissues or organs to exhibit neuronal activity. Cellular memory involves multiple components, including protein, RNA, and DNA modifications, but the

precise mechanisms sustaining cellular energy activity remain unclear. Improving cellular energy metabolism—such as enhancing mitochondrial function or supplying metabolic substrates—can significantly inhibit myofibroblast activation. This strategy holds promise for reversing heart failure by promoting the transdifferentiation of myofibroblasts into fibroblasts, thereby restoring cardiac function[20].

To investigate the epigenetic regulation mechanisms of energy metabolism, we conducted a multimodal analysis of the single-cell RNA sequencing (scRNA-seq) and single-cell epigenomic data of NSCs provided by Kremer et al.[2] First, active neural stem cells (aNSCs) and quiescent neural stem cells type 2 (qNSC2) were extracted based on labels provided by Kremer et al.[2], and the scRNA-seq dataset (GSE209656) was analyzed using TOGGLE (Figure 13).

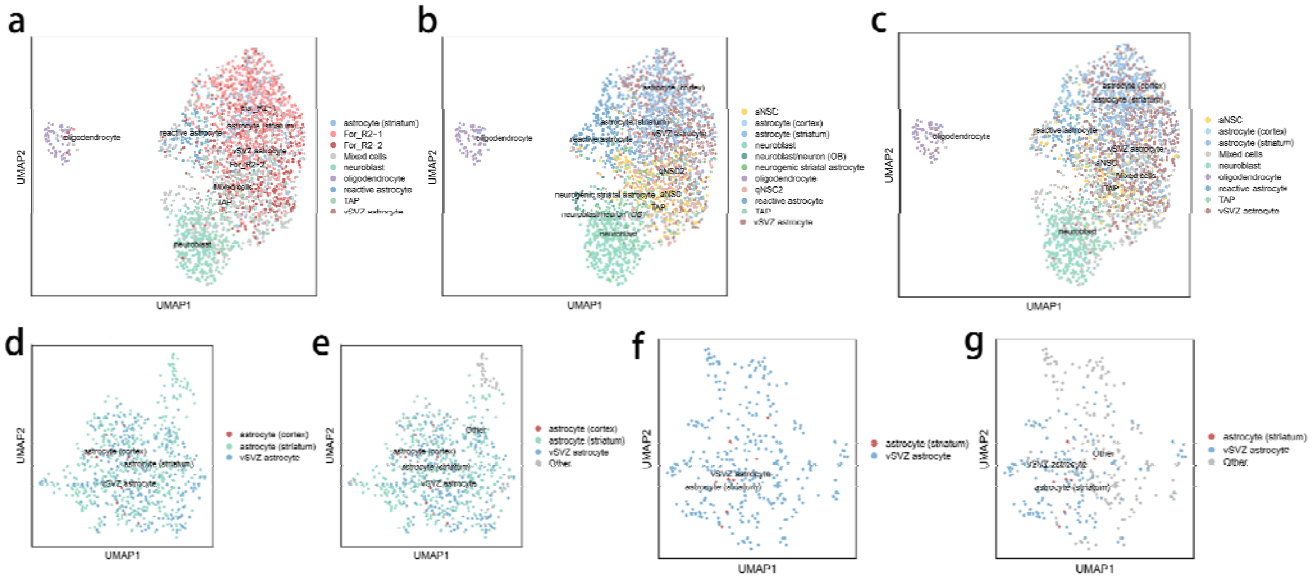


Figure 13 A: Preliminary prediction results indicating that For_R2-1 and For_R2-2 are highly similar and difficult-to-classify groups. B: True labels. C: Results after completing classifications D and F. D: Predicted labels for For_R2-1. E: True labels for For_R2-1. F: Predicted labels for For_R2-2. G: True labels for For_R2-2.

Next, we divided the cells (Figure 14a and 14b) into two groups based on the median mitochondrial RNA expression levels: a high mitochondrial RNA expression group (High-MT) and a low mitochondrial RNA expression group (Low-MT) (Figure 14d and 14e). The analysis revealed that NeuG04 predominantly consisted of metabolically active cells, mainly active neural stem cells (aNSCs), whereas NeuG01 was mainly composed of metabolically inactive cells, primarily quiescent neural stem cells (qNSCs) (Figure 14c and 14f). The remaining cell groups were distributed between these two

categories, possibly representing transitional metabolic states. To identify differentially methylated regions between the two groups, we performed two-sided Wilcoxon rank-sum tests on VMRs covered by at least six cells in each group. Regions with a Benjamini–Hochberg corrected P-value less than 0.05 were labeled as low methylation regions (LMRs) or highly accessible chromatin regions (high ACC). LMRs overlapping gene body regions were annotated to their corresponding genes.

Since NeuG01 predominantly consists of aNSC cells and NeuG04 mainly comprises qNSC2 cells, we conclude the following: (1) During the transition from qNSC2 to aNSC, local DNA demethylation occurs in key genomic regions associated with cell proliferation and energy metabolism (Figure 14g), accompanied by significantly increased chromatin accessibility (Figure 14h). (2) The proportion of cells exhibiting high mitochondrial gene expression (High-MT) within the aNSC-dominated group is significantly higher than in the qNSC2 group (Figure 14c, 14d, and 14e), further indicating enhanced metabolic activity in aNSCs. In summary, we propose that during the transition from qNSC2 to aNSC, local DNA demethylation and chromatin opening activate the expression of genes related to cell proliferation and energy metabolism, gradually enhancing metabolic activity. These epigenetic modifications reflect a difference in "cellular memory" between the two cell types, potentially serving as the underlying mechanism driving their distinct mitochondrial metabolic states.

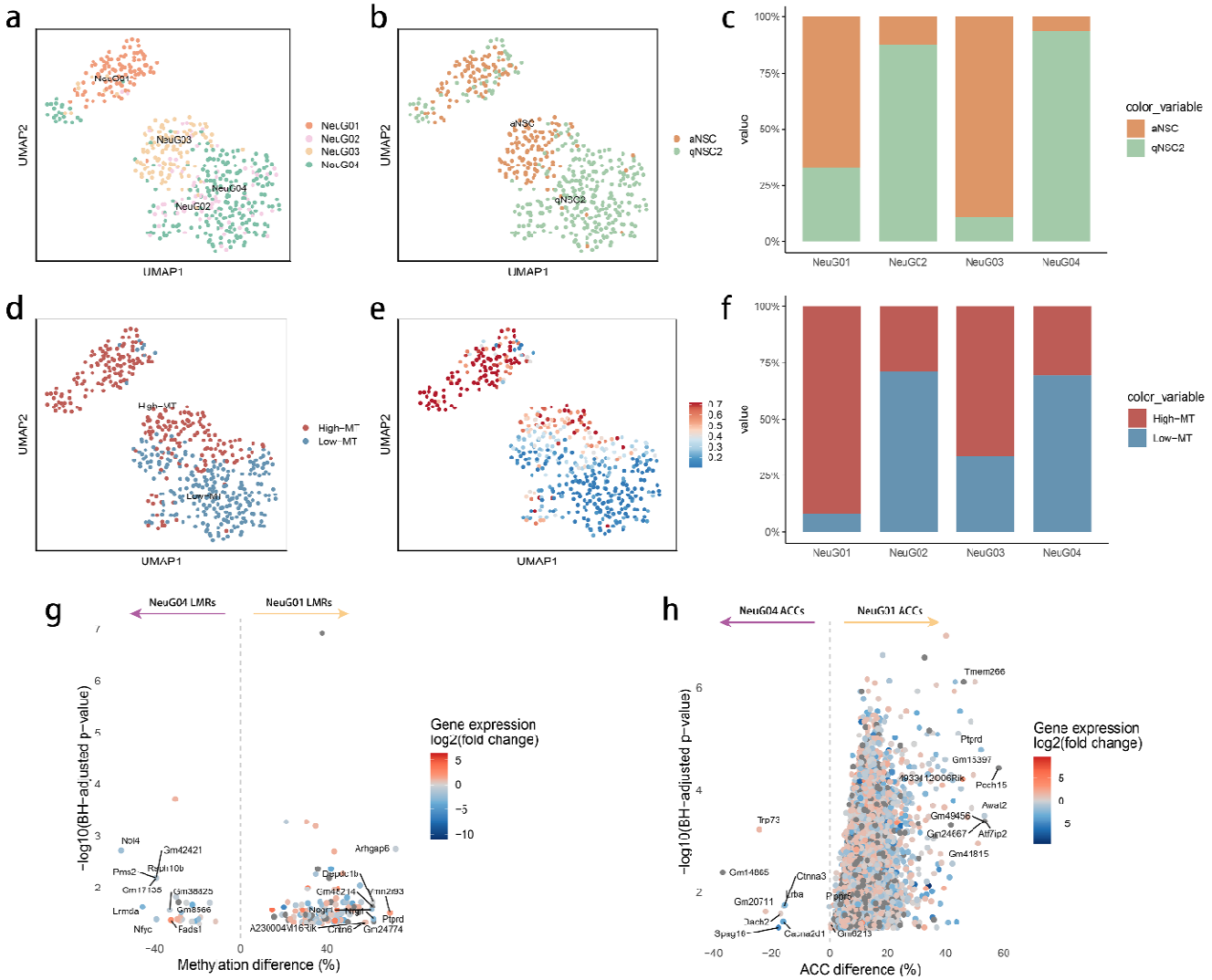


Figure 14 Results of epigenetic analysis of aNSCs and qNSCs. a: Cell populations after TOGGLE analysis; b: Labels provided by Kremer et al.; c: The proportion of aNSC and qNSC2 in each cell group; d: Cells are divided into High-MT and Low-MT according to the expression of mitochondrial genes; e: Mitochondrial 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

3.4 Other test results and findings

In this study, we applied the TOGGLE algorithm to the functional classification of cell-cell communication[21~25] and RNA pathways. Due to space constraints, detailed methodological descriptions are provided in the appendix, and only a summary is presented here. In the e-cigarette aerosol exposure model (GSE183614), TOGGLE grouped differentially expressed genes and conducted enrichment

analyses, identifying that Groups 2 through 5 exhibited significant immune and inflammatory responses across biological processes and KEGG pathways (e.g., TNF signaling and cell cycle). This effect was particularly pronounced in younger mice, indicating that e-cigarette exposure has more severe impacts on immature cardiac tissue. In the myocardial infarction (MI) model (GSE253768), TOGGLE integrated single-nucleus RNA sequencing (snRNA-seq) data with bulk RNA sequencing data to identify seven functional subpopulations of fibroblasts and myofibroblasts (FibR1-G1 to G7), with the MI group predominantly composed of FibR1-G5 through G7. Further analysis of cell-cell ligand-receptor interactions revealed differences in their communication patterns with immune cells. TOGGLE subsequently mapped bulk RNA-seq data onto the snRNA-seq space, discovering that the bulk RNA-seq signals primarily originated from FibR2-G6, a population enriched in pathways related to cell migration, adhesion, angiogenesis, and ECM-receptor interactions, suggesting its role as a key cell type driving tissue-level changes. In summary, TOGGLE not only precisely dissects cellular heterogeneity but also integrates single-nucleus and bulk RNA sequencing data across platforms to uncover critical signaling pathways and functional cellular populations. These results are detailed in the Supplementary Appendix 1&2.

4. Discussion

This study proposes TOGGLE, an unsupervised learning framework that does not require manual labeling. Combining graph diffusion, self-supervised mechanisms, and a BERT-like deep encoding network, TOGGLE achieves high-precision identification of cell functional states and fate stages from single-cell transcriptomic data. Additionally, TOGGLE supports integrated analysis of single-cell transcriptomic and epigenomic datasets. In neural stem cell (NSC) data, cells were divided into High-MT and Low-MT groups based on mitochondrial RNA expression levels. The analysis revealed that the transition from qNSC to aNSC is accompanied by increased chromatin accessibility and DNA demethylation, elucidating the epigenetic memory mechanism underlying metabolic activity differences. Compared to mainstream algorithms, TOGGLE has advantages in distinguishing cell functional states, tracking cell fate trajectories, integrating cross-modal data, eliminating the need for pre-training, and enabling multifunctional analyses within similar cell populations (see Table 2 for details). It thus holds broad applicability in regenerative medicine, immunotherapy, responses to viral infections, and cancer research.

While TOGGLE demonstrates robust performance in integrating transcriptomic and epigenomic data and tracing cell fate transitions, it also has several limitations. Firstly, it does not yet incorporate post-transcriptional regulatory signals such as RNA methylation (e.g., m6A), which play critical roles in

modulating gene expression. Secondly, the algorithm's ability to resolve rare or minor cell subpopulations remains to be further improved, as these subtypes often carry important biological functions in complex tissues. Lastly, TOGGLE currently does not include the analysis of post-translational protein modifications (e.g., phosphorylation, ubiquitination), which are key regulators of signaling and cell functional states. Future work could aim to integrate these additional regulatory layers to provide a more comprehensive picture of cell state and functionality.

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.

Key points：

（1） Except for immune cells, mature cells typically do not undergo clonal expansion and exhibit a high degree of similarity. Our tool is capable of classifying cellular functions in different contexts.

（2） We developed an innovative Graph Diffusion Functional Map, which significantly reduces noise, thereby providing clearer displays of RNA functional groupings and enabling the capture of more subtle functional differences in high-dimensional data.

（3） We have implemented a novel tracking technique capable of distinguishing between early, intermediate, and late stages of cellular programming.

（4） In the testing dataset, it was found that local DNA demethylation activates the expression of genes associated with cell proliferation and energy metabolism, thereby regulating the activation of quiescent neurons. Neuronal ferroptosis in ischemic stroke may be linked to the high expression of ferroptosis driver genes such as Ctsb, Mtdh, Ndrg1, Smad7, Cd82, and Acsl1.

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*+, Zhouweiyu Chen+, Kailin Yang+* and Eric Jiang. 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 Kailin Yang was 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. Junpeng Chen proposed a complete set of biological problems and performed the analysis, with assistance from Kailin Yang, Kaiqing Liu, Tianda Sun, Yibing Nong, Tao Yuan, 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. “

. . ( ) 1828521

, .”

s.

, : ,

, and .

References

1. Nakamura A, Sakai S, Taketomi Y, et al. PLA2G2E-mediated lipid metabolism triggers brain-autonomous neural repair after ischemic stroke. *Neuron*. 2023;111(19):2995-3010.e9. doi:10.1016/j.neuron.2023.06.024

2. Kremer LPM, Cerrizuela S, El-Sammak H, et al. DNA methylation controls stemness of astrocytes in health and ischaemia. *Nature*. 2024;634(8033):415-423. doi:10.1038/s41586-024-07898-9

3. Schiebinger G, Shu J, Tabaka M, et al. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*. 2019;176(4):928-943.e22. doi:10.1016/j.cell.2019.01.006

4. Wang SW, Herriges MJ, Hurley K, Kotton DN, Klein AM. CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nat Biotechnol*. 2022;40(7):1066-1074. doi:10.1038/s41587-022-01209-1

5. Weinreb C, Rodriguez-Fraticelli A, Camargo FD, Klein AM. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*. 2020;367(6479):eaaw3381. doi:10.1126/science.aaw3381

6. Biddy BA, Kong W, Kamimoto K, et al. Single-cell mapping of lineage and identity in direct reprogramming. *Nature*. 2018;564(7735):219-224. doi:10.1038/s41586-018-0744-4

7. Cerrizuela S, Kaya O, Kremer LPM, et al. High-throughput scNMT protocol for multiomics profiling of single cells from mouse brain and pancreatic organoids. *STAR Protoc*. 2022;3(3):101555. doi:10.1016/j.xpro.2022.101555

8. Neczypor EW, Saldaña TA, Mears MJ, et al. e-Cigarette Aerosol Reduces Left Ventricular Function in Adolescent Mice. *Circulation*. 2022;145(11):868-870. doi:10.1161/CIRCULATIONAHA.121.057613

9. Jiang J, Ye X, Kong Y, et al. scLTdb: a comprehensive single-cell lineage tracing database. *Nucleic Acids Research*. Published online October 29, 2024:gkae913. doi:10.1093/nar/gkae913

10. Jiang J, Li J, Huang S, et al. CACIMAR: cross-species analysis of cell identities, markers, regulations, and interactions using single-cell RNA sequencing data. *Briefings in Bioinformatics*. 2024;25(4):bbae283. doi:10.1093/bib/bbae283

11. Jiang J, Lyu P, Li J, et al. IReNA: Integrated regulatory network analysis of single-cell transcriptomes and chromatin accessibility profiles. *iScience*. 2022;25(11):105359. doi:10.1016/j.isci.2022.105359

12. 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*. 2024;7(1):1205. doi:10.1038/s42003-024-06833-y

13. Sledge IJ, Príncipe JC. Partitioning Relational Matrices of Similarities or Dissimilarities Using the Value of Information. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ; 2018:2416-2420. doi:10.1109/ICASSP.2018.8462453

14. Prasad N, Yang K, Uhler C. Optimal Transport using GANs for Lineage Tracing. Published online 2020. doi:10.48550/ARXIV.2007.12098

15. Sadria M, Zhang A, Bader GD. Deep Lineage: Single-Cell Lineage Tracing and Fate Inference Using Deep Learning. Published online April 26, 2024:2024.04.25.591126. doi:10.1101/2024.04.25.591126

16. Sadria M, Swaroop V. Discovering governing equations of biological systems through representation learning and sparse model discovery. *NAR Genom Bioinform*. 2025;7(2):lqaf048. doi:10.1093/nargab/lqaf048

17. Zhou N, Yuan X, Du Q, et al. FerrDb V2: update of the manually curated database of ferroptosis regulators and ferroptosis-disease associations. *Nucleic Acids Research*. 2023;51(D1):D571-D582. doi:10.1093/nar/gkac935

18. Kanehisa M, Sato Y, Kawashima M. KEGG mapping tools for uncovering hidden features in biological data. *Protein Sci*. 2022;31(1):47-53. doi:10.1002/pro.4172

19. Lan B, Ge JW, Cheng SW, et al. Extract of Naotaifang, a compound Chinese herbal medicine, protects neuron ferroptosis induced by acute cerebral ischemia in rats. *J Integr Med*. 2020;18(4):344-350. doi:10.1016/j.joim.2020.01.008

20. Younesi FS, Miller AE, Barker TH, Rossi FMV, Hinz B. Fibroblast and myofibroblast activation in normal tissue repair and fibrosis. *Nat Rev Mol Cell Biol*. 2024;25(8):617-638. doi:10.1038/s41580-024-00716-0

21. Peng, L., Liu, L., Huang, L., Bai, Z., Chen, M., & Chen, X. (2025). Predicting cell–cell communication by combining heterogeneous ensemble deep learning and weighted geometric mean. Applied Soft Computing, 112839.

22. Peng, L., Liao, B., Zhu, W., Li, Z., & Li, K. (2015). Predicting drug–target interactions with multi-information fusion. IEEE journal of biomedical and health informatics, 21(2), 561-572.

23. Peng, L., Gao, P., Xiong, W., Li, Z., & Chen, X. (2024). Identifying potential ligand–receptor interactions based on gradient boosted neural network and interpretable boosting machine for intercellular communication analysis. Computers in Biology and Medicine, 171, 108110.

24. Peng, L., Tan, J., Xiong, W., Zhang, L., Wang, Z., Yuan, R., ... & Chen, X. (2023). Deciphering ligand–receptor-mediated intercellular communication based on ensemble deep learning and the joint scoring strategy from single-cell transcriptomic data. Computers in Biology and Medicine, 163, 107137.

25. Peng, L., Xiong, W., Han, C., Li, Z., & Chen, X. (2023). CellDialog: a computational framework for ligand-receptor-mediated cell-cell communication analysis. IEEE Journal of Biomedical and Health Informatics, 28(1), 580-591.