Supplementary Information for

TFcomb identifies transcription factor combinations for cellular reprogramming based on single-cell multi-omics data

Chen Li[1], Sijie Chen[1], Yixin Chen[1], Haiyang Bian[1], Minsheng Hao[1], Lei Wei[1,*] and Xuegong Zhang[1,2]

[1] Ministry of Education Key Laboratory of Bioinformatics, Bioinformatics Division at the Beijing National Research Center for Information Science and Technology, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing 100084, China

[2] Center for Synthetic and Systems Biology, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China

[*] Corresponding author: weilei92@tsinghua.edu.cn

Supplementary Notes

Supplementary Note 1: Graph attention networks enhance gene regulatory signals

We adopted the GRN benchmark in CellOracle to validate the improvement of GAT model on the original GRN. Five tissues available both in the Tabula Muris scRNA-seq dataset (Schaum et al. 2018) and mouse sci-ATAC-seq atlas data (Cusanovich et al. 2018) were selected as the ground-truth datasets: heart, kidney, liver, lung and spleen. The ground-truth GRNs were curated using 1,298 chromatin immunoprecipitation followed by sequencing (ChIP–seq) datasets for totally 80 regulatory factors across these tissues (Kamimoto et al. 2023). The TF number of each ground-truth GRN ranges from 7 to 44, and the TF-target connection number ranges from 340 to 33,247. The scATAC-seq data were provided by mouse sci-ATAC-seq atlas (Cusanovich et al. 2018), and 13 samples across 5 tissues in the scRNA-seq dataset were utilized to conduct the GRN benchmark.

For each sample, we used CellOracle to calculate a primary GRN, and then GAT was applied to enhance the GRN. As for the GAT training, we performed a ten-fold cross-validation approach to improve the quality of the recovering TF-target links (Methods). As shown in Fig. S1A, the GAT prediction model gives the prediction result with a high area under the precision-recall curve (AUPRC) score on the test sets across all the 13 samples, which indicates the model robustly learns the regulatory relations between TFs and targets for each fold. For example, in sample Liver_2 and Spleen_0, the prediction model achieves AUPRC over 0.980 across all folds. The fold-average AUROC score ranges from 0.917 (Kidney_2) to 0.995 (Lung_0) across samples, which indicates GAT model is highly effective in capturing the regulatory signals of different tissues.

To further evaluate the GAT-recovered TF-target links, we calculated the accuracy (ACC) of the predicted links by comparing them to the ground-truth GRN for each TF across different samples. Specifically, for each TF, a recovered link was considered correct if it was present in the ground-truth GRN. We aggregated the data across all 13 samples in Fig. S1A, resulting in a total of 109 TFs and their targets as the ground truth. For each TF, we randomly sampled unconnected links whose number is equal to the number of predicted links as the control group. Among these, GAT outperforms the random method in 50 TFs, shows comparable performance in 30 TFs, and performs worse in 29 TFs (Fig. S1B). Moreover, the overall performance metrics of GAT were significantly higher than those of

the random method (one-side Wilcoxon test p-value = 9.6×10^{-4}, Fig. S1C). These results highlight the ability of GAT to recover biologically meaningful regulatory links.

Additionally, we illustrated the recovered links of a specific TF *Creb1*, a kind of cyclic adenosine monophosphate (cAMP) responsive element modulator, in sample Kidney_0 (Fig. S1D). 8 of 12 links (red lines) are in the ground-truth kidney GRN, and *Arpc4* and *Mthrf* are demonstrated to be targets of *Creb1* in JASPAR predicted transcription factor targets dataset (Rouillard et al. 2016; Castro-Mondragon et al. 2022). All the results demonstrate that the GAT module captures the regulatory relations and enhances the regulatory signals of the primary GRN.

Moreover, to evaluate the effectiveness of the GAT encoder in our GRN scenario, we compared it to three other encoders: MLP, GCN (Kipf and Welling 2016), and GraphSAGE (Hamilton et al. 2017). Both GCN and GraphSAGE are GNNs that lack an attention mechanism (Fig. S2). For each base GRN in the benchmarking cell states, we trained the models using a tenfold scheme. The average accuracy on the test sets was used to represent the performance. Overall, GAT (ACC=0.975) outperformed MLP (ACC=0.971), GCN (ACC=0.973), and GraphSAGE (ACC=0.974). While MLP provided relatively satisfactory results, we found GAT to be a more effective and robust encoder for learning node features. We further replaced the MLP prediction module with a dot-product-based prediction mechanism, which resulted in a significant performance drop compared to the GAT model (Fig. S2). This suggests that the GAT mechanism captures complex relationships beyond the simple dot product.

Supplementary Note 2: Parameter settings

Recovery ratio

To better understand the impact of the recovery ratio, we conducted experiments to explore how the number of recovered links and identification performance vary with different recovery ratios. We tested recovery ratios of 0.5%, 1%, 5%, 10%, 20%, 30%, 40%, and 50%. As shown in Fig. S14, the number of recovered links exhibited a linear relationship with the recovery ratio. At a recovery ratio of 30%, the GAT module has already recovered more links than the original GRN contained. Regarding identification performance, the recovery ratio of 10% achieved the best result. Overall, performance initially increased with the recovery ratio and then declined.

We hope to balance the number of recovered links with identification performance. On the one hand, we aim for the GAT module to recover an appropriate number of links without disrupting the base GRN structure. On the other hand, we seek to recover links that enhance identification performance. Based on this balance, we selected a recovery ratio of 5% as the default value.

Weight of different propagation steps

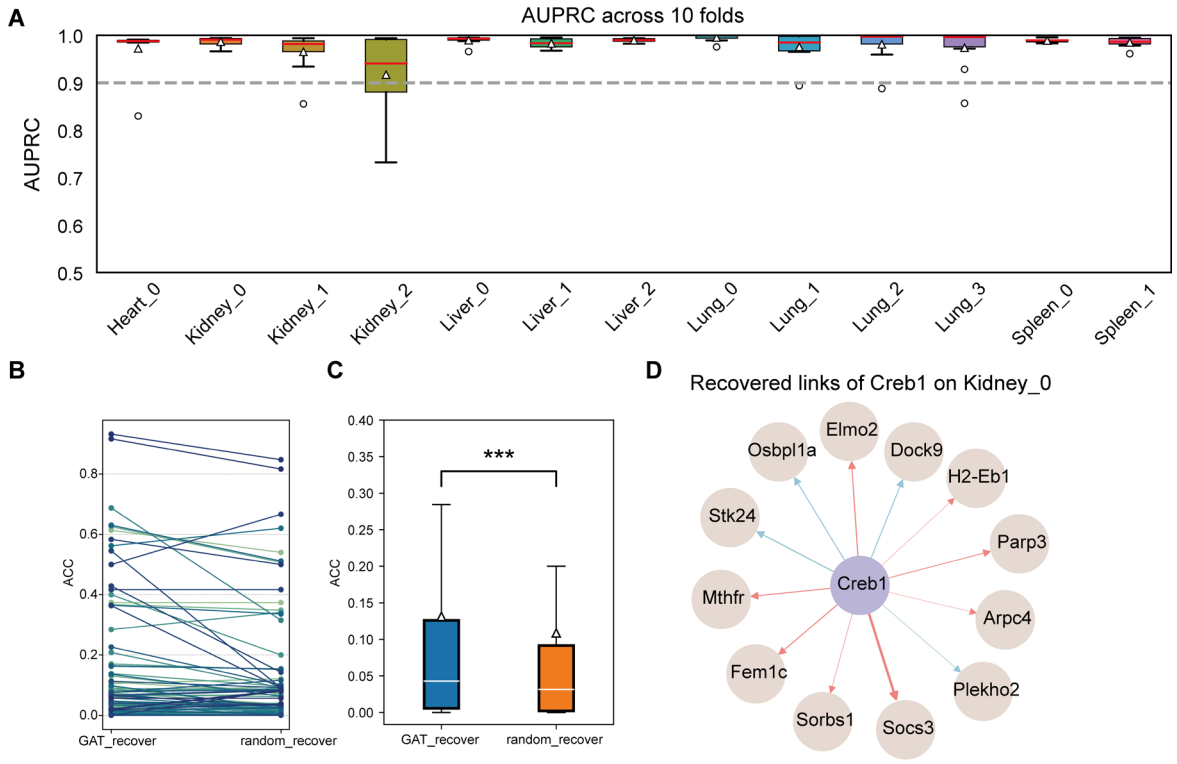
Regarding the propagation step parameters, they represent the weights of different propagation steps. To analyze their impact, we conducted an experiment testing various propagation step settings and compared the performance across the benchmarking states. The tested settings were:

1. $\lambda_1 = 1, \lambda_2 = 0, \lambda_3 = 0$;
2. $\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = 0$;
3. $\lambda_1 = 0, \lambda_2 = 0, \lambda_3 = 1$;
4. $\lambda_1 = 0.6, \lambda_2 = 0.2, \lambda_3 = 0.2$.

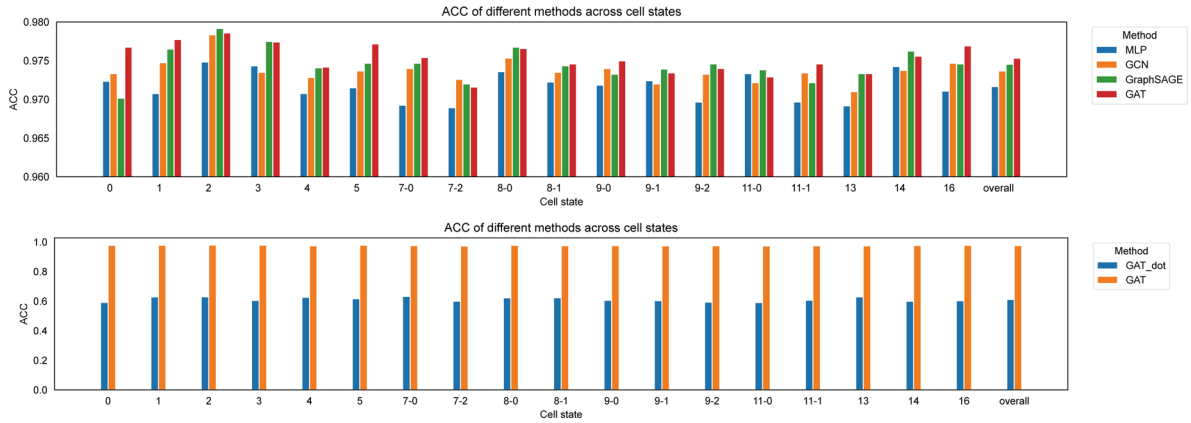
The results demonstrate that model performance is significantly influenced by the propagation steps across different cell states (Fig. S15). For instance, in state 7-2, setting the two-step propagation weight (λ_2) to 1 achieved the best performance, while in state 8-0, the three-step propagation (λ_3) performed best. In contrast, in some states either two-step propagation or three-step propagation decline the model performance (state 0, 5, 8-1, and 11-0). These findings suggest that different cell

states may require distinct propagation step settings. For example, in certain states, reprogramming TFs may require more propagation steps to exert their influence. To achieve a relatively balanced and robust result, we selected $\lambda_1 = 0.6, \lambda_2 = 0.2, \lambda_3 = 0.2$ as the default propagation parameters.

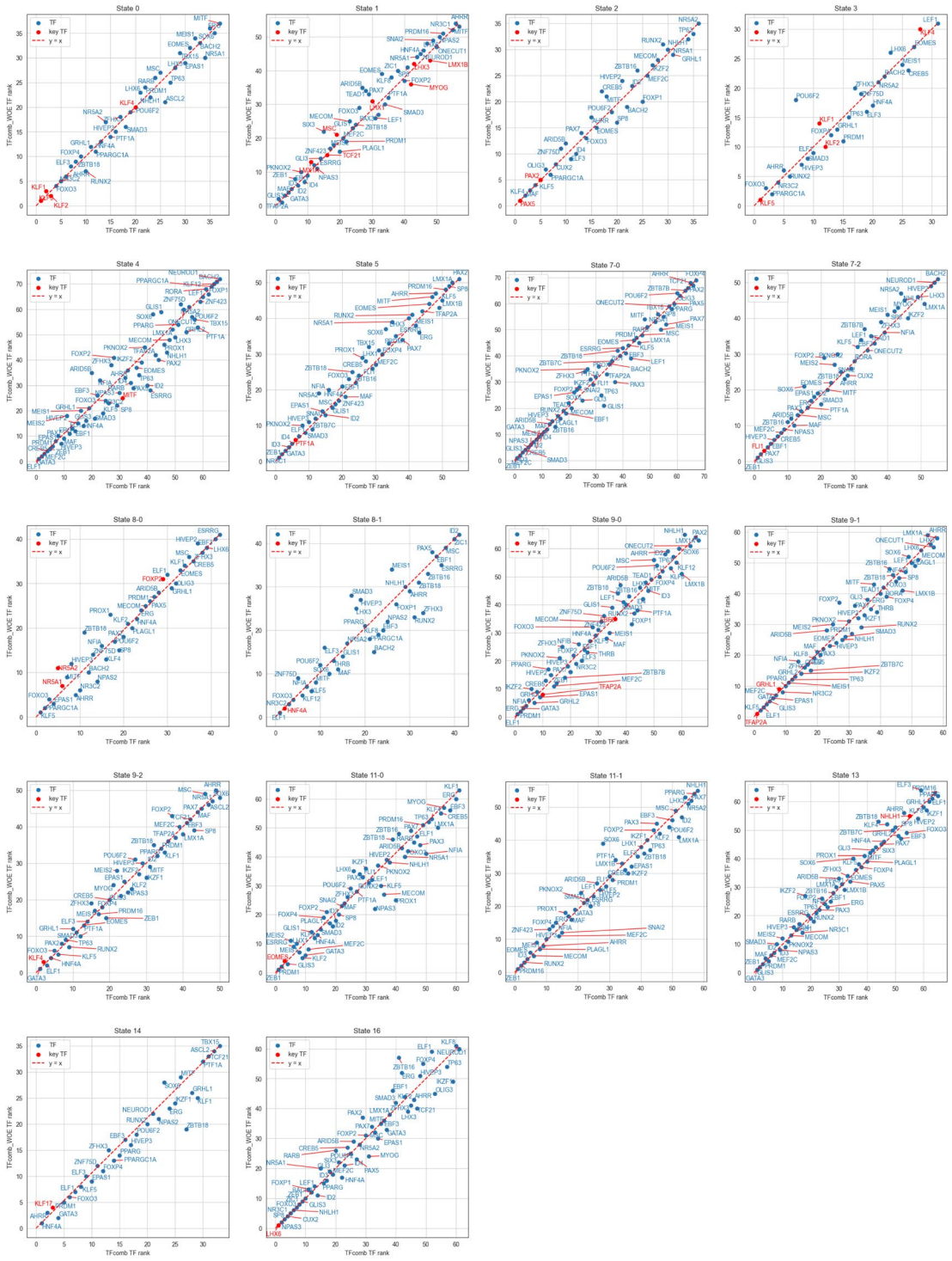
Supplementary Figures



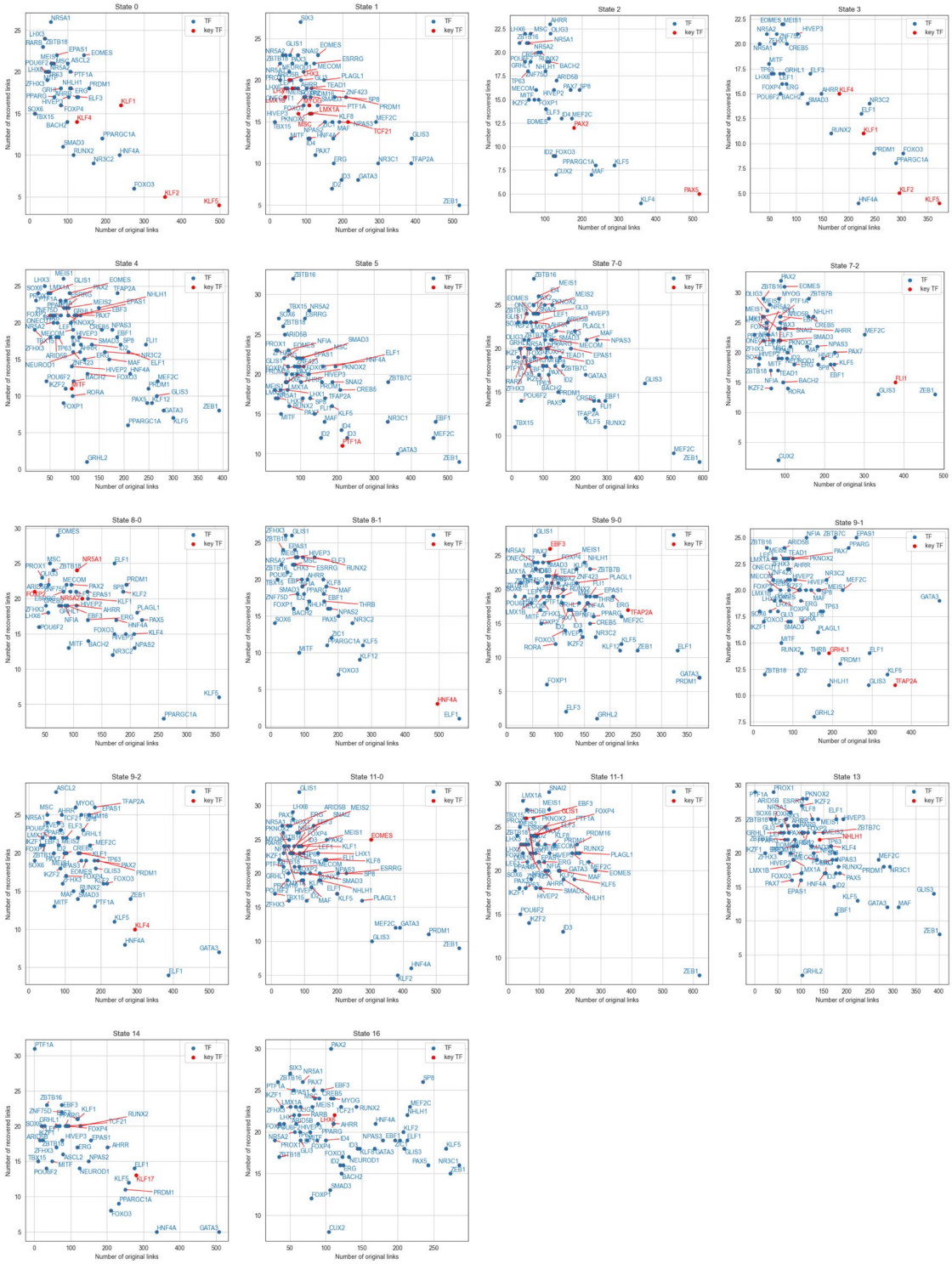
Supplementary Fig. 1. Graph attention network (GAT) module enhances the regulatory signal of original GRN. (A) Boxplots of area under the receiver operating characteristic (AUROC) on the test set across 13 samples. The box plots indicate the medians (centerlines), means (triangles), first and third quartiles (bounds of boxes). **(B-C)** Performance of recovering TF-target links on 109 TFs across 13 samples. The significance level of one-side Wilcoxon test *p*-value was shown in (C). ***, *p*-value < 0.001. **(D)** GAT recovered network of *Creb1*. Ground-truth links are annotated in red.



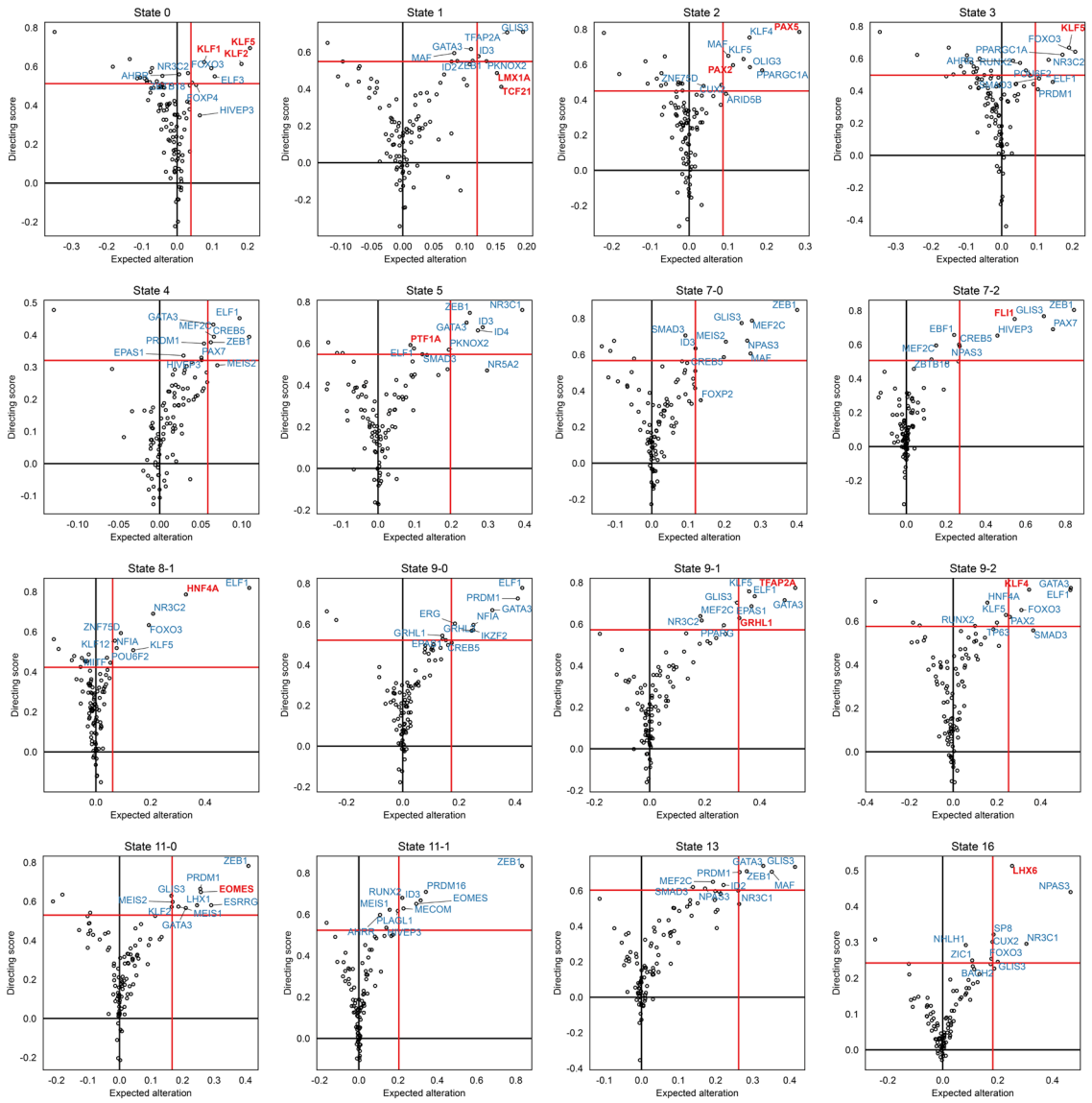
Supplementary Fig. 2. Comparison of GAT to other graph neural networks. Top: Performance comparison of MLP, GCN, GraphSAGE, and GAT. Bottom: Performance comparison of GAT with and without the dot product predictor.



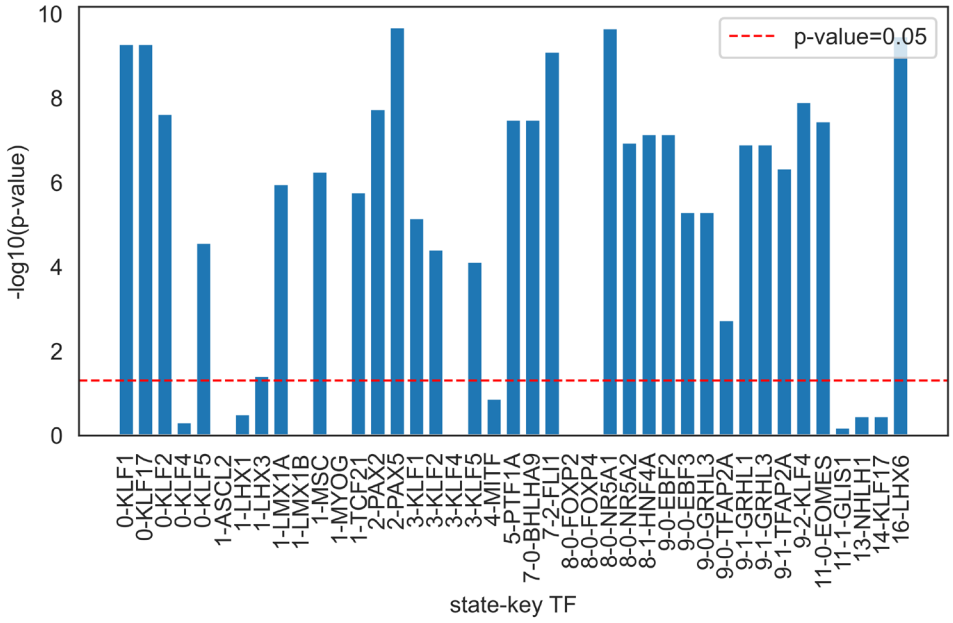
Supplementary Fig. 3. Comparison of TF rankings between TFcomb and TFcomb_WOE. The x-axis represents the TF rankings from TFcomb, and the y-axis represents those from TFcomb_WOE. Key TFs are highlighted in red. Points above the reference line (the dashed red line) show improved rankings with TFcomb relative to TFcomb_WOE.



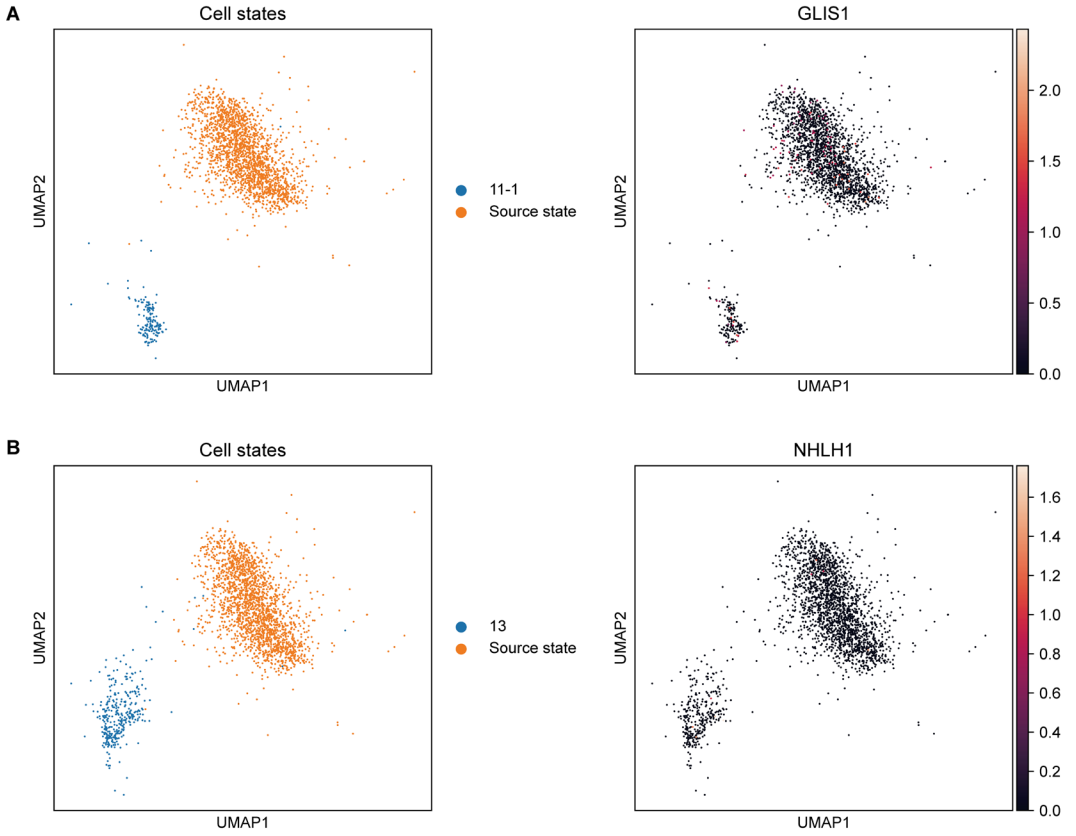
Supplementary Fig. 4. The comparison of the counts of original links and recovered links among all TFs.



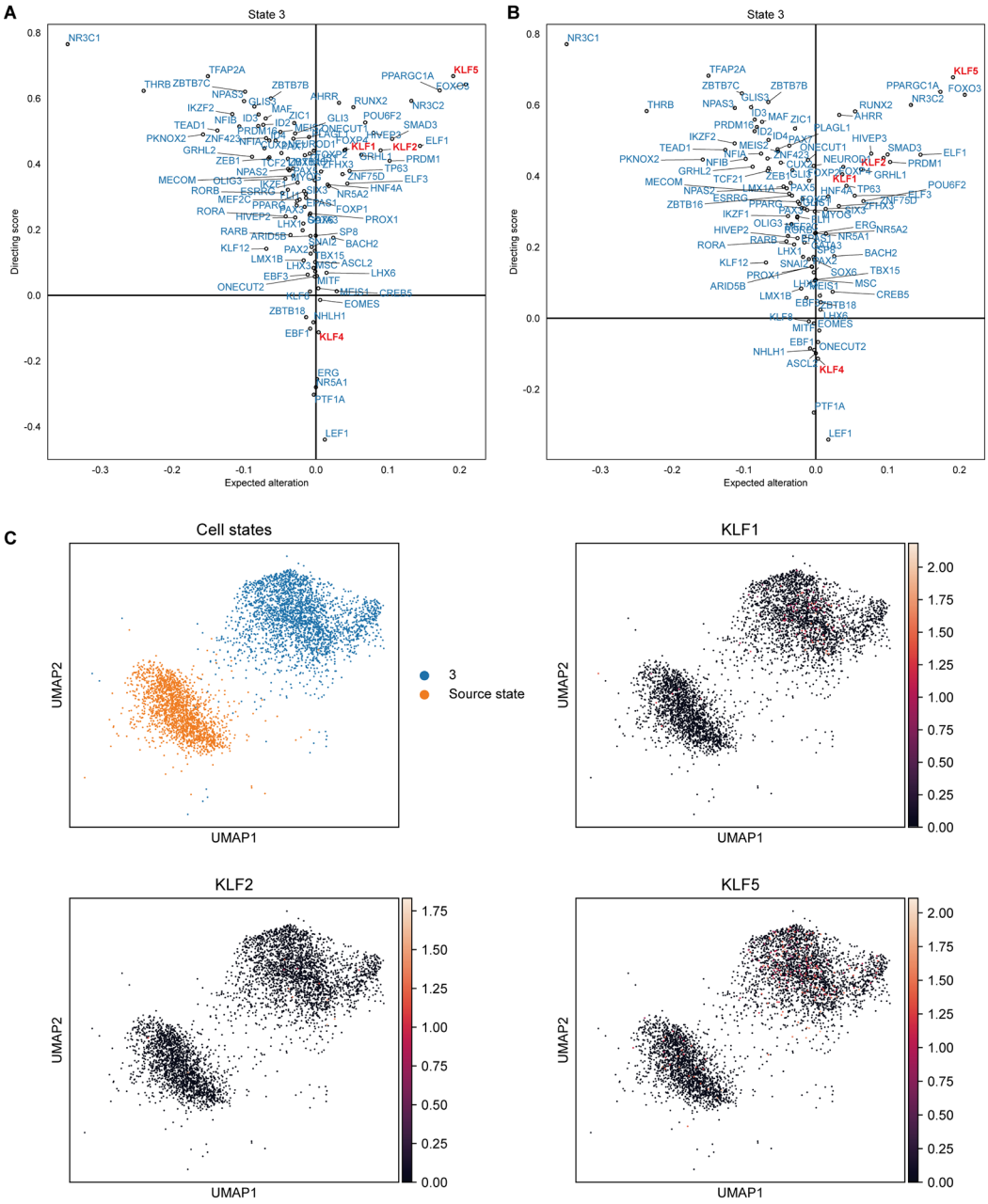
Supplementary Fig. 5. TFcomb TF identification plot on other target cell states. Red lines are the quantile thresholds to filter 10 TFs. Key TFs are annotated in red.



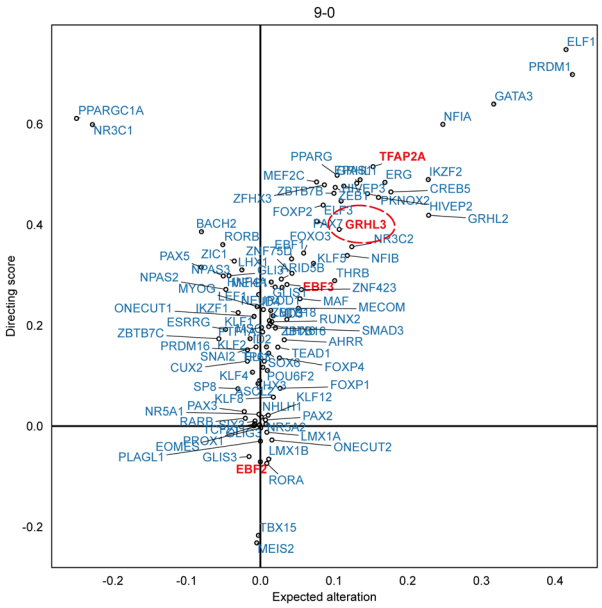
Supplementary Fig. 6. Significance level of the one-side *t*-statistic test on whether the directing score of the key TF is significantly higher across all states.



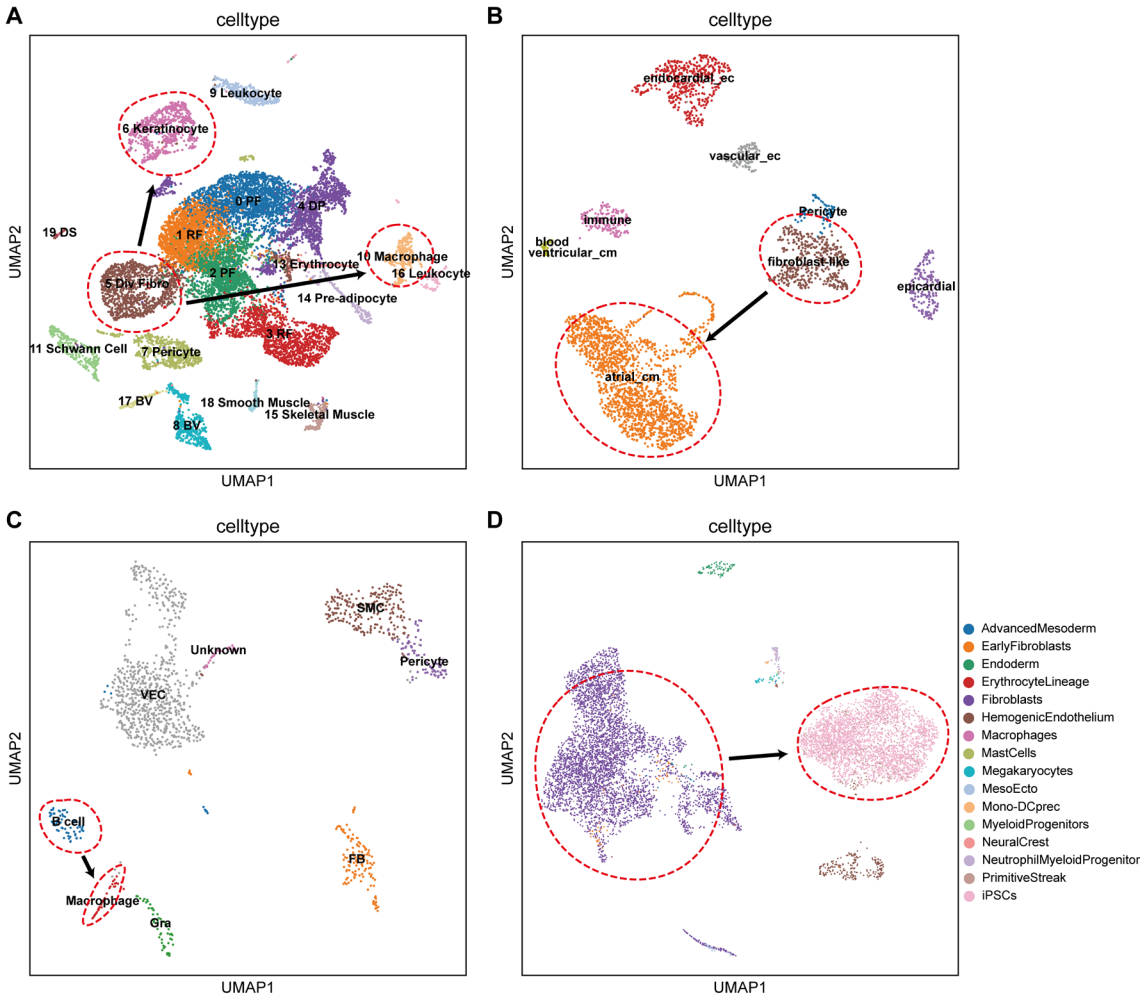
Supplementary Fig. 7. (A) The UMAP visualization of the source state and the target state 11-1. The normalized expression of key TF *GLIS1* is shown on the UMAP embeddings. **(B)** The UMAP visualization of the source state and the target state 13. The normalized expression of key TF *NHLH1* is shown on the UMAP embeddings.



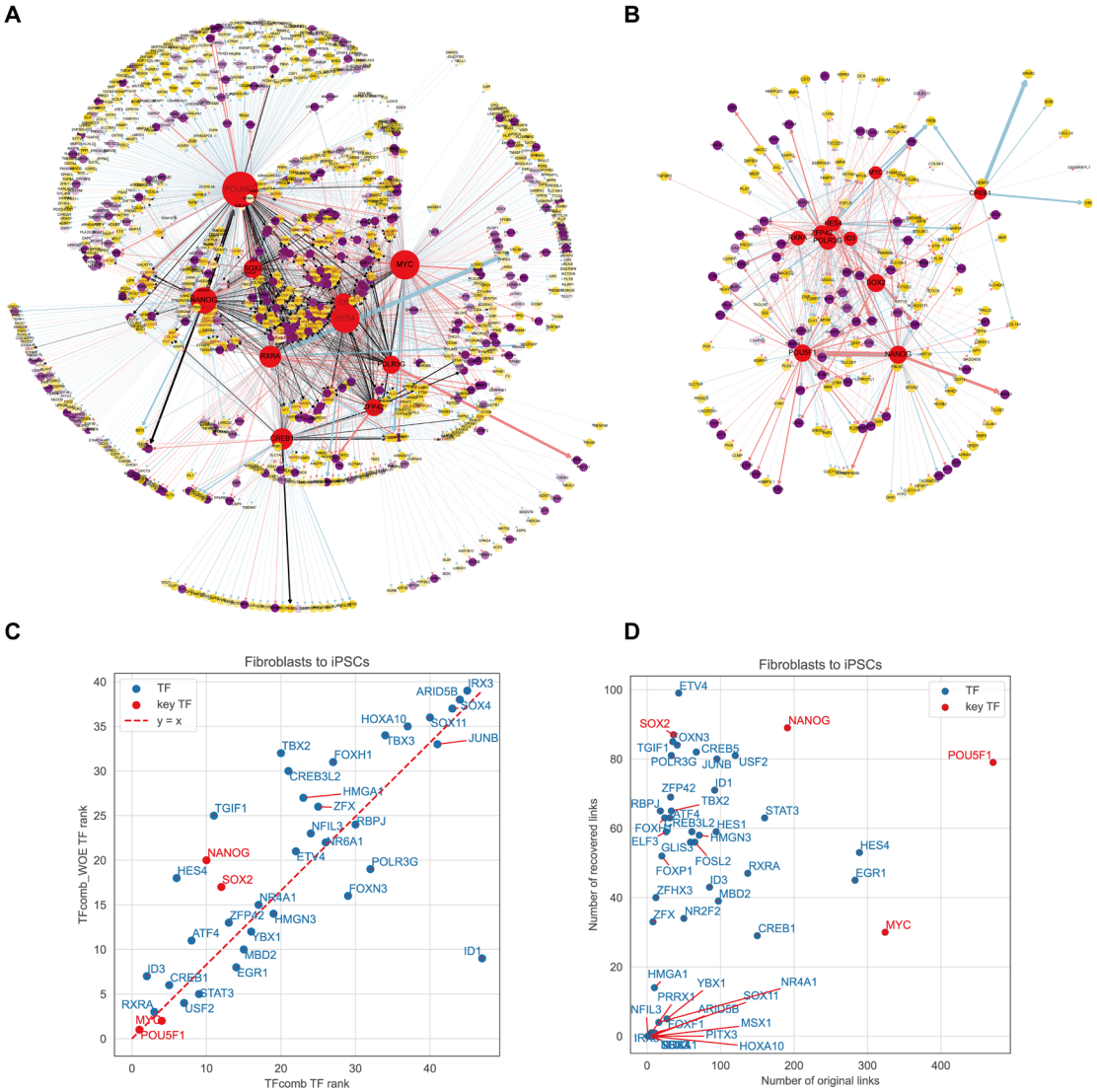
Supplementary Fig. 8. (A) TFcomb TF identification plot on target state 3. Key TFs are annotated in red. **(B)** TFcomb_WOE TF identification plot on target state 3. Key TFs are annotated in red. **(C)** The UMAP visualization of the source state and the target state 3. The normalized expression of key TF *KLF1*, *KLF2* and *KLF5* is shown on the UMAP embeddings.



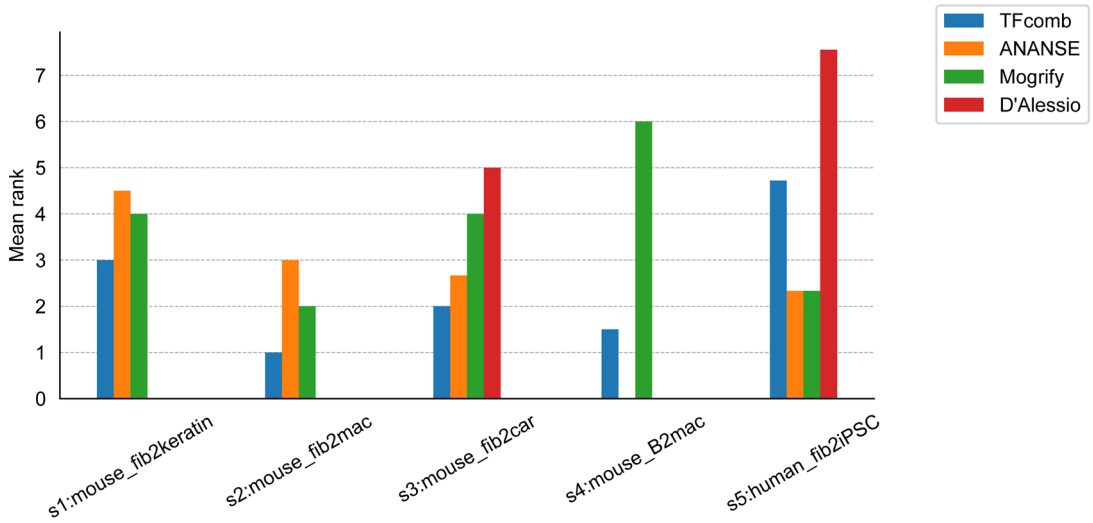
Supplementary Fig. 9. TFcomb TF identification plot on target state 9-0. The knowledge of NicheNet database is added to supplement the TF-target links. Key TFs are annotated in red. The added key TF *GRHL* is annotated in a red circle.



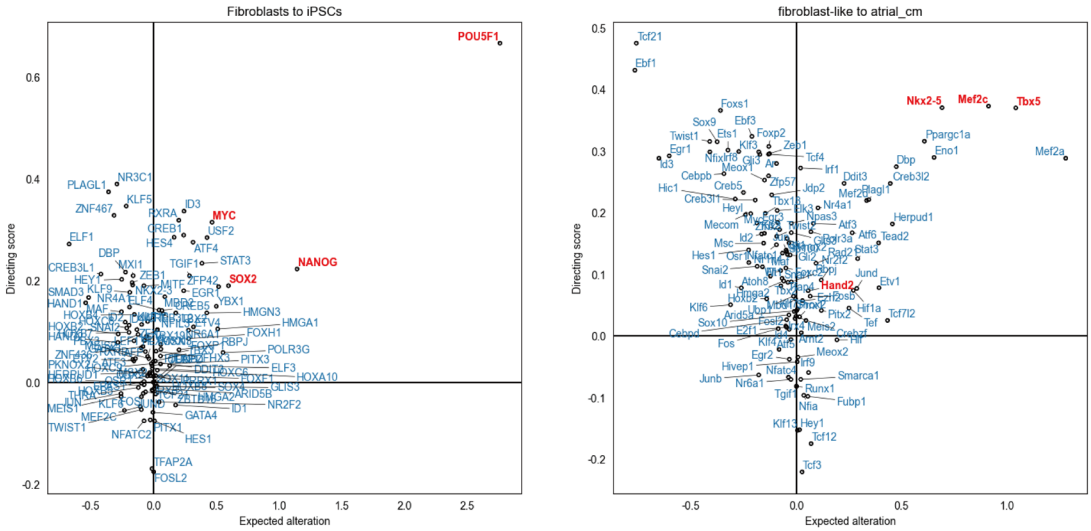
Supplementary Fig. 10. UMAP visualization of single-cell datasets for reprogramming cases. The dashed red circles represent source or target cell states. The black arrows point from source cell states to target cell states. **(A)** Fibroblasts to keratinocytes or macrophages in mouse. **(B)** Fibroblasts to cardiomyocytes in mouse. **(C)** B cells to macrophages in mouse. **(D)** Fibroblasts to induced pluripotent stem cells (iPSCs) in human.



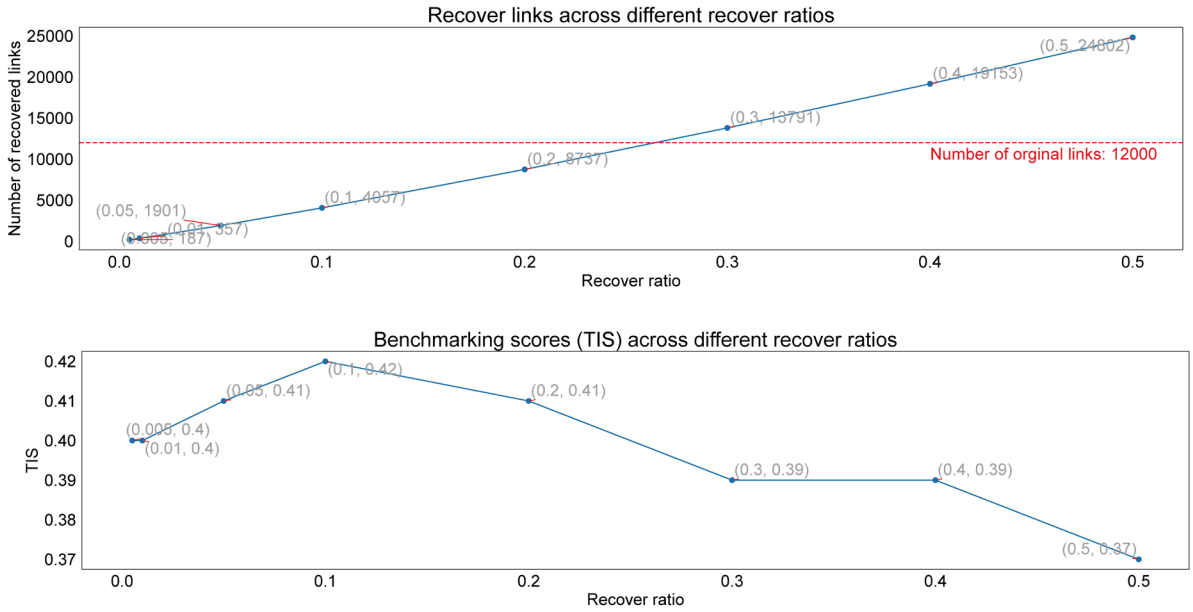
Supplementary Fig. 11. Illustration of GAT recovered links of human fibroblasts reprogramming to iPSCs. (A) Visualization of the GRN with the top 10 TFs and their targets. Recovered links are annotated in black. TFs are annotated in red. The pruple nodes and yellow nodes represent the target genes differentially expressed in the target state and the source state, respectively. The red lines and blue lines denote the positive links and negative links, respectively. (B) Visualization of GRN with top 10 TFs and their recovered links. (C) The ranking comparison of TFcomb and TFcomb_WOE. (D) The count comparison of original links and recovered links.



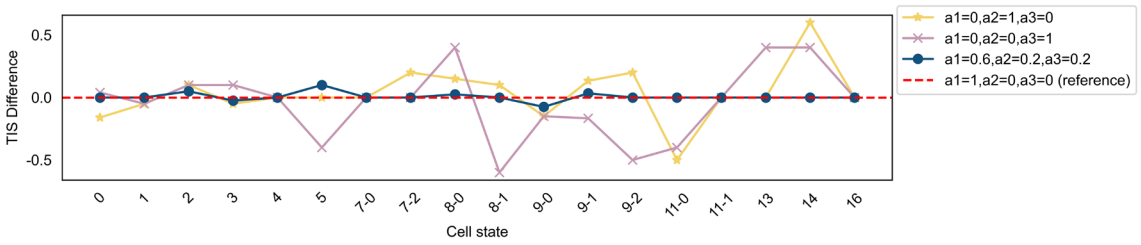
Supplementary Fig. 12. TF identification comparison by mean rank across different reprogramming cases.



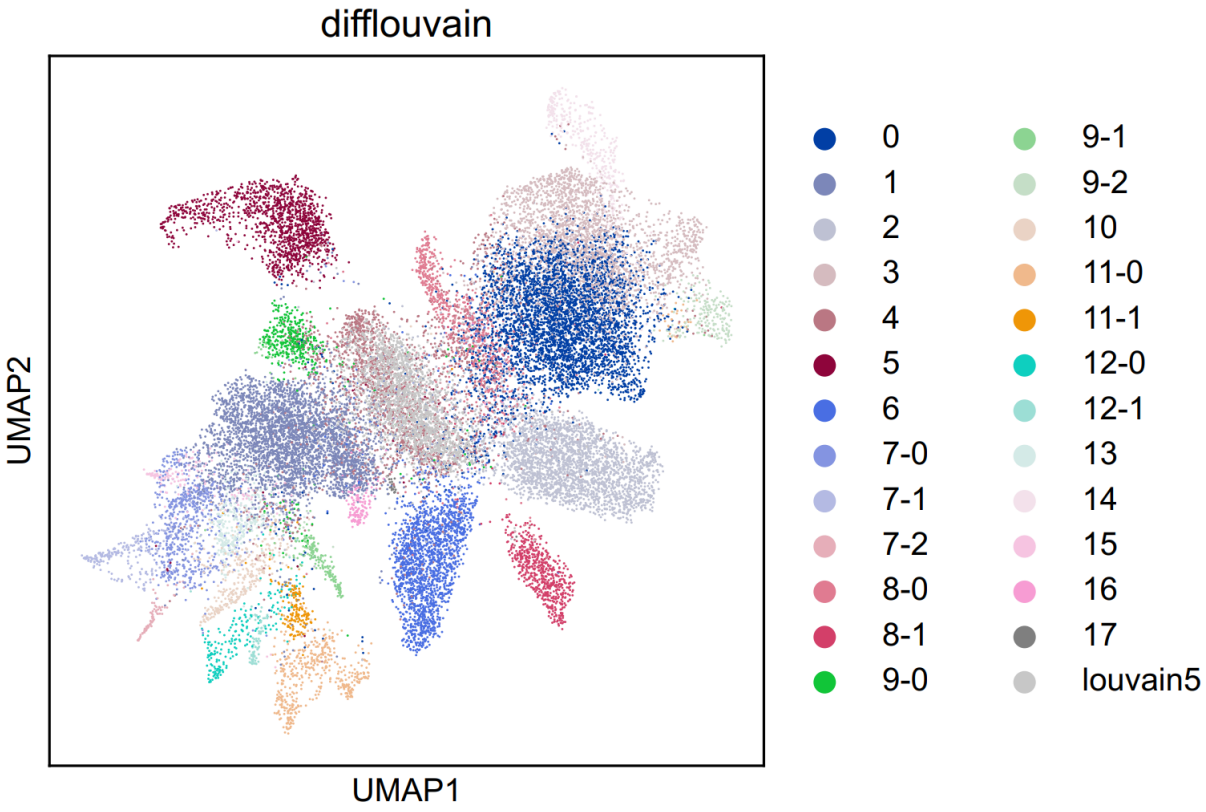
Supplementary Fig. 13. Left: TFcomb TF identification plot of human fibroblasts to iPSCs. Right: TFcomb TF identification plot of mouse fibroblasts to cardiomyocytes.



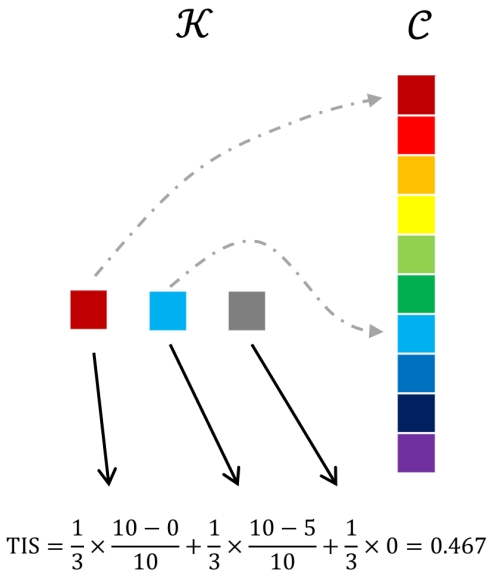
Supplementary Fig. 14. The number of recovered links and identification performance at different recovery ratios. Top: Line plot showing the relationship between the number of recovered links and the recovery ratio. Bottom: Line plot illustrating the variation of TIS with the recovery ratio.



Supplementary Fig. 15. TIS variation across different parameter settings for propagation steps.



Supplementary Fig. 16. UMAP visualization of 2,000 sampled cells and 25 original target cells. The 'louvain5' is the cluster of sampled cells.



Supplementary Fig. 17. The illustration of TIS calculation.

References

Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Lemma RB, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Perez NM et al. 2022. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **50**: D165-D173.

Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS et al. 2018. A Single-Cell Atlas of *In Vivo* Mammalian Chromatin Accessibility. *Cell* **174**: 1309-+.

Hamilton WL, Ying R, Leskovec J. 2017. Inductive Representation Learning on Large Graphs. In *31st Annual Conference on Neural Information Processing Systems (NIPS)*, Vol 30, Long Beach, CA.

Kamimoto K, Stringa B, Hoffmann CM, Jindal K, Solnica-Krezel L, Morris SA. 2023. Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**: 742-+.

Kipf TN, Welling MJapa. 2016. Semi-supervised classification with graph convolutional networks. *arXiv*.

Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, Ma'ayan AJD. 2016. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**.

Schaum N Karkanias J Neff NF May AP Quake SR Wyss-Coray T Darmanis S Batson J Botvinnik O Chen MB et al. 2018. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. *Nature* **562**: 367-+.