CytoFlow supplementary information

1 Data preparation

To comprehensively demonstrate CytoFlow's application potential, we applied CytoFlow to three datasets.

1.1 Bulk transcriptome data from Saccharomyces cerevisiae

As a proof-of-concept experiment, we incorporated two types of networks to reconstruct the signal transduction network in *Saccharomyces cerevisiae*: the PPI network and the gene co-expression network derived from microarray data. The baseline network was the mitogen-activated protein kinase ($MAPK$) pheromone response pathway [1] documented in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [2, 3, 4].

We first constructed the PPI network using public data from the Database of Interacting Proteins (DIP) [5] and StringDB [6]. We incorporated the network edges of the DIP *Saccharomyces cerevisiae* PPI core network (version 20080708, with 17,414 interactions documented among 4,901 genes) with the interaction confidence scores of the physical interactions documented in StringDB, resulting in a network with 4,901 nodes and 10,840 edges. Then, we removed any gene that was at least two steps away from the pheromone response pathway to reduce computational cost, resulting in the final weighted network with 115 nodes and 230 edges as inputs for CytoFlow (**Table S3**).

Next, to construct the gene co-expression network, we utilized *Saccharomyces cerevisiae* gene expression profiling array data from the Gene Expression Omnibus (GEO) with the accession number GSE8895 [7]. After removing genes not documented in StringDB [6], the expression profiles of 5,513 genes were extracted from 12 samples. Then, we built the gene co-expression network by calculating the pairwise Pearson correlation among all genes. To obtain a high-confidence prior network, we only kept the edges in the raw DIP PPI network and removed any edge with a negative correlation value [5]. Finally, we removed distal genes from the pheromone response pathway using the same procedure as in PPI, resulting in a network with 135 nodes and 300 edges (**Table S3**).

1.2 scRNA-seq data preprocessing

We also applied CytoFlow to the gene co-expression networks constructed from two public scRNA-seq datasets from post-mortem brains [8] and peripheral blood mononuclear cells [9].

Expression data preprocessing

- *Human prefrontal cortex (PFC)*: we downloaded the raw reads from three prefrontal cortex (GSE216270) [8] and pre-processed the data using the default parameters mentioned in Duan et al. The data contains 25,555 cells from seven major cell types in the human brain, with the number of cells from each cell type listed in **Table S5**.
- *Human peripheral blood mononuclear cells (PBMCs)*: Similarly, we downloaded the public PBMC 3k and 10k multi-omics datasets [9] from 10x Genomics and extracted the scRNA-seq part, which contains 3,000 genes from 14,266 cells (from an initial 180,488 barcodes and 36,601 genes). We applied the public available cell type annotation, whose composition is listed in **Table S5**. To perform a robust and reliable analysis, we removed small cell types and kept only five distinct major cell types: *CD14*-positive monocytes (*CD14 Mono*), *CD4*-positive central memory T-cells (*CD14 TCM*), *CD16*-positive monocytes (*CD16 Mono*), natural killer cells (*NK*), and intermediate B-cells (*B intermediate*).

Gene metainformation processing In both studies, we only used the 19683 protein coding genes downloaded from the GENCODE GRCh38 release [10]. Human receptors were downloaded from LRdb in SingleCellSignalR [11], and TFs were downloaded from RegNetwork [12]. After intersecting with the protein-coding genes, we get 745 receptors and 1,442 TFs. All other protein-coding genes are treated as potential intermediate nodes. We calculated the cell-type-specific average log10-normalized expression for every gene included in this study as input to the model.

1.3 Construction of cell-type-specific gene co-expression networks

Here we describe our pipeline to construct cell-type-specific gene co-expression networks for scRNA-seq data inspired by [13, 14, 15, 16].

Gene selection To obtain a robust network from sparse data, we selected genes that are highly expressed (top average expression) and robustly expressed (present in at least 5% of the cells) for each cell type. The number of genes selected was 3,000 for PFC and 500 for PBMC. To demonstrate the cell-type specificity of the reconstructed signal transduction networks, we then selected five cell-type-specific receptors and TFs for each cell type (**Table S2**). Specifically, in a given cell type, we selected the top five expressed receptors (TFs)

that are more highly expressed than in any other cell types. The expression ratio threshold was 1.5 for PFC and 1 for PBMC. We removed all other receptors and TFs and obtained the final gene list, with which we constructed the cell-type-specific gene co-expression networks.

Metacell aggregation We constructed *metacells* by combining cells with similar transcriptomic profiles to overcome the sparsity of scRNA-seq data. Cells of the same type were aggregated into *metacells* by averaging signals from similar cells, identified from a k-nearest neighbor (KNN) graph (default $k = 100$) constructed from the first 20 PCA dimensions computed by the standard Seurat data processing pipeline [17]. We then identified the maximum number of *metacells* such that the overlap ratio (the number of common neighbors divided by k) of any two *metacells* was less than 0.8. Specifically, we traversed all *metacells* in a random order, greedily selecting a *metacell* unless its overlap ratio with any selected *metacell* was over 0.8. Finally, the selected *metacells* were normalized with a library size of 10,000.

Cell-type-specific gene co-expression network construction We constructed the cell-type-specific gene co-expression networks using the *metacells* by calculating the pairwise Pearson correlations among all genes. To reduce computational cost, we only retained the top 50 edges connected to each node with at least 0.5 edge weight for the PFC network (a node was removed if it had no retained edges). For the PBMC network, the numbers were 20 edges and 0.4 edge weight, respectively.

2 Experiment details

2.1 Tuning the node and edge penalty parameters for CytoFlow

We proposed a novel parameter tuning criterion for running CytoFlow to construct signal transduction networks in a self-supervised manner. We observed an interesting phenomenon regarding the relationship between the parameters and the absolute penalty $\lambda_n \sum_i w_i + \lambda_e \sum_{(i,j)\in E} \widehat{f_{(i,j)}}$ applied in the objective function. The penalty starts low when both λ_n and λ_e (because the multipliers are low), rises as the parameters increase, and drops again when either λ_n or λ_e is high (because the flows and node weights are small). When the input network has a moderate density, the highest penalty always occurs when the total flow and output network connectivity are both at reasonable levels. We claim that a high penalty biologically implies that the parameters represent a reasonable tradeoff between message passing and energy consumption, indicating that the organism makes full use of the energy consumed.

Thus, we tuned the parameters by maximizing the penalty through a grid search. We first ran CytoFlow over a two-dimensional parameter grid ($\lambda_n, \lambda_e = 0.001, 0.02$ to 0.30 with step size 0.02), recording each run's optimal objective,

absolute penalty, and total flow. For the experiments performed on *Saccharomyces cerevisiae*, we utilized the parameter combination that maximized the flow among all parameter settings with at least 90% of the maximum penalty. Note that we recommend not setting either parameter to zero, but rather a small positive value. When the node penalty is zero, the output network may contain redundant nodes whose weights are non-zero but do not transmit any flow. When the edge penalty is zero, there's a chance that the flow network will contain directed loops, which is not biologically informative.

We applied this parameter tuning pipeline to the *Saccharomyces cerevisiae* networks. Parameter choices can be found in **Table S4**. To fairly compare all pairwise flows, we set all parameters of all receptor-TF pairs to 0.01 in our second and third experiments.

2.2 Running established methods

We tuned the parameter of Zhao et. al. according to the pipeline described in the original paper [18]. Ren et. al. did not provide a detailed parameter tuning scheme, so we selected the best parameter mentioned in their paper [19]. Parameter choices in every run can be found in **Table S4**. We then applied thresholds to the optimal networks for evaluation. The parameters of Zhao et al. are designed to be binary in the optimal solution, so we removed all nodes and edges with weights of at most 0.5. Ren et al. outputs weighted networks, so we kept only edges with weights above 0.001 and removed all nodes not connected to any edges.

References

[1] L. Bardwell, "A walk-through of the yeast mating pheromone response pathway," *Peptides*, vol. 26, no. 2, pp. 339–50, 2005.

[2] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, 2000.

[3] M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, and M. Ishiguro-Watanabe, "Kegg for taxonomy-based analysis of pathways and genomes," *Nucleic Acids Res*, vol. 51, no. D1, pp. D587–D592, 2023.

[4] M. Kanehisa, "Toward understanding the origin and evolution of cellular organisms," *Protein Sci*, vol. 28, no. 11, pp. 1947–1951, 2019.

[5] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg, "Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Res*, vol. 30, no. 1, pp. 303–5, 2002.

[6] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J.

Jensen, and C. von Mering, "The string database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest," *Nucleic Acids Res*, vol. 51, no. D1, pp. D638–D646, 2023.

[7] P. Daran-Lapujade, M. L. Jansen, J. M. Daran, W. van Gulik, J. H. de Winde, and J. T. Pronk, "Role of transcriptional regulation in controlling fluxes in central carbon metabolism of saccharomyces cerevisiae. a chemostat culture study," *J Biol Chem*, vol. 279, no. 10, pp. 9125–38, 2004.

[8] Z. Duan, Y. Dai, A. Hwang, C. Lee, K. Xie, C. Xiao, M. Xu, M. J. Girgenti, and J. Zhang, "iherd: an integrative hierarchical graph representation learning framework to quantify network changes and prioritize risk genes in disease," *PLoS Comput Biol*, vol. 19, no. 9, p. e1011444, 2023.

[9] J. Zhong, M. Qiu, Y. Meng, P. Wang, S. Chen, and L. Wang, "Single-cell multi-omics sequencing reveals the immunological disturbance underlying stat3-v637m hyper-ige syndrome," *Int Immunopharmacol*, vol. 122, p. 110624, 2023.

[10] A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J. E. Loveland, J. M. Mudge, C. Sisu, J. C. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, C. Boix, S. Carbonell Sala, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. Garcia Giron, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, K. L. Howe, T. Hunt, O. G. Izuogu, R. Johnson, F. J. Martin, L. Martinez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, F. C. Riera, M. Ruffier, B. M. Schmitt, E. Stapleton, M. M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, M. Y. Wolf, J. Xu, Y. T. Yang, A. Yates, D. Zerbino, Y. Zhang, J. S. Choudhary, M. Gerstein, R. Guigo, T. J. P. Hubbard, M. Kellis, B. Paten, M. L. Tress, and P. Flicek, "Gencode 2021," *Nucleic Acids Res*, vol. 49, no. D1, pp. D916–D923, 2021.

[11] S. Cabello-Aguilar, M. Alame, F. Kon-Sun-Tack, C. Fau, M. Lacroix, and J. Colinge, "Singlecellsignalr: inference of intercellular networks from single-cell transcriptomics," *Nucleic Acids Res*, vol. 48, no. 10, p. e55, 2020.

[12] Z. P. Liu, C. Wu, H. Miao, and H. Wu, "Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse," *Database (Oxford)*, vol. 2015, 2015.

[13] L. Zhang, J. Zhang, and Q. Nie, "Direct-net: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data," *Sci Adv*, vol. 8, no. 22, p. eabl7393, 2022.

[14] S. Morabito, E. Miyoshi, N. Michael, S. Shahin, A. C. Martini, E. Head, J. Silva, K. Leavy, M. Perez-Rosendahl, and V. Swarup, "Single-nucleus chromatin accessibility and transcriptomic characterization of alzheimer's disease," *Nat Genet*, vol. 53, no. 8, pp. 1143–1155, 2021.

[15] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, p. 559, 2008.

[16] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Stat Appl Genet Mol Biol*, vol. 4, p. Article17, 2005.

[17] Y. Hao, T. Stuart, M. H. Kowalski, S. Choudhary, P. Hoffman, A. Hartman, A. Srivastava, G. Molla, S. Madad, C. Fernandez-Granda, and R. Satija, "Dictionary learning for integrative, multimodal and scalable single-cell analysis," *Nat Biotechnol*, 2023.

[18] X. M. Zhao, R. S. Wang, L. Chen, and K. Aihara, "Uncovering signal transduction networks from high-throughput data by integer linear programming," *Nucleic Acids Res*, vol. 36, no. 9, p. e48, 2008.

[19] X. W. Ren and X. S. Zhang, "A linear programming model based on network flow for pathway inference," *Journal of Systems Science & Complexity*, vol. 23, no. 5, pp. 971–977, 2010.