iMIRACLE: an Iterative Multi-View Graph Neural Network to Model Intercellular Gene Regulation from Spatial Transcriptomic Data

Ziheng Duan
zihend1@uci.edu
University of California, Irvine
Irvine, CA, USA

Siwei Xu
s.xu@uci.edu
University of California, Irvine
Irvine, CA, USA

Cheyu Lee
cheyul1@uci.edu
University of California, Irvine
Irvine, CA, USA

Dylan Riffle
driffle@uci.edu
University of California, Irvine
Irvine, CA, USA

Jing Zhang*
zhang.jing@uci.edu
University of California, Irvine
Irvine, CA, USA

Abstract

Spatial transcriptomics has transformed genomic research by measuring spatially resolved gene expressions, allowing us to investigate how cells adapt to their microenvironment via modulating their expressed genes. This essential process usually starts from cell-cell communication (CCC) via ligand-receptor (LR) interaction, leading to regulatory changes within the receiver cell. However, few methods were developed to connect them to provide biological insights into intercellular regulation. To fill this gap, we propose iMiracle, an iterative multi-view graph neural network that models each cell's intercellular regulation with three key features. Firstly, iMiracle integrates inter- and intra-cellular networks to jointly estimate *cell-type*- and *micro-environment*-driven gene expressions. Optionally, it allows prior knowledge of intra-cellular networks as pre-structured masks to maintain biological relevance. Secondly, iMiracle employs iterative learning to overcome the sparsity of spatial transcriptomic data and gradually fill in the missing edges in the CCC network. Thirdly, iMiracle infers a cell-specific ligand-gene regulatory score based on the contributions of different LR pairs to interpret inter-cellular regulation. We applied iMiracle to nine simulated and eight real datasets from three sequencing platforms and demonstrated that iMiracle consistently outperformed ten methods in gene expression imputation and four methods in regulatory score inference. Lastly, we developed iMiracle as an open-source software and anticipate that it can be a powerful tool in decoding the complexities of inter-cellular transcriptional regulation.

CCS Concepts

• **Applied computing** → **Bioinformatics**; • **Computing methodologies** → **Machine learning approaches**.

*Jing Zhang is the corresponding author.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679574

Keywords

inter-cellular gene regulation; cell–cell communications; spatial transcriptomics; graph neural networks

ACM Reference Format:
Ziheng Duan, Siwei Xu, Cheyu Lee, Dylan Riffle, and Jing Zhang. 2024. iMIRACLE: an Iterative Multi-View Graph Neural Network to Model Intercellular Gene Regulation from Spatial Transcriptomic Data. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3627673.3679574

1 Introduction

In eukaryotic organisms, precise spatial and temporal regulation of transcription is crucial for a range of fundamental biological processes, from development to adaptation to disease progression [6, 15, 16, 39–41, 57]. Thanks to concerted community efforts and technological advancements, there has been a remarkable leap over the past decades in our understanding of transcription regulation within individual cells [4, 7, 17, 23, 56, 59]. Thus, it has opened avenues for therapeutic strategies targeting specific transcriptional pathways and mechanisms [28]. While the current use of transcriptional technologies is promising, cells live in an organized combination of extracellular matrix, cells, and interstitial fluid that jointly influence gene expression [11, 46]. Aberrations in such intercellular communications within this spatial context may disrupt gene expression profiles, ultimately leading to cellular changes and pathogenic outcomes [24]. Despite its importance, the exploration of inter-cellular communication and its downstream impacts on transcriptional regulation remains underdeveloped. This gap limits our ability to fully understand multi-cellular functions and their implications for health and disease, highlighting an urgent need for new computational efforts.

To bridge this gap, we propose a novel, iterative multiview graph neural network (GNN) model named iMiracle to investigate intercellular transcriptional regulation for each cell. This model is designed with ***three distinct features*** to tackle the current challenges. First, iMiracle integrates inter- and intra-cellular networks for accurate expression imputation using ligand-receptor interactions with neighboring cells. Optionally, it allows users to include prior knowledge of intra-cellular networks, such as protein-protein interaction network (PPI) and gene regulatory network (GRN), as

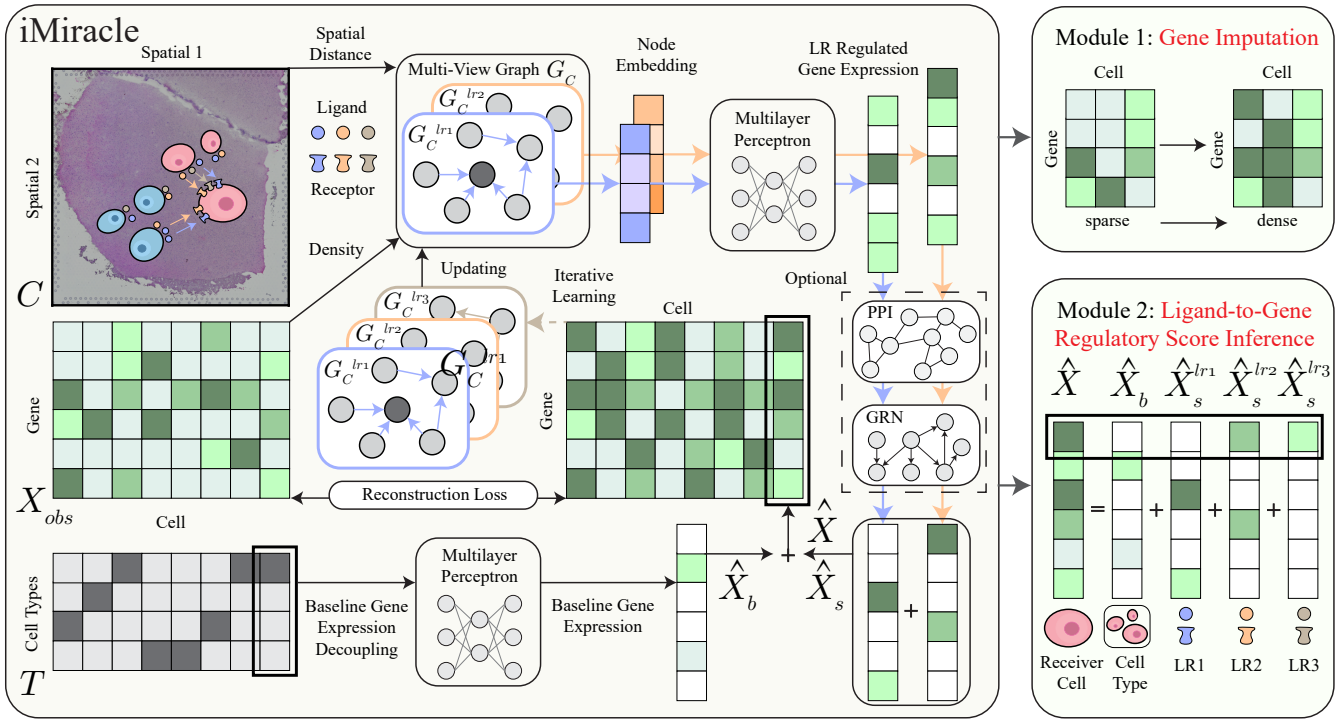

Figure 1: Overview of iMiracle. iMiracle initiates with a sparse cell-by-gene matrix $X_{obs} \in \mathbb{R}^{n \times m}$ (n cells and m genes), spatial coordinates $C \in \mathbb{R}^{n \times 2}$, and cell type information $T \in \mathbb{R}^{n \times t}$ (t cell types). It constructs multi-view cell-cell communication networks G_C to model various ligand-receptor interactions. Node embeddings for each cell are generated per view through a graph neural network. A multilayer perceptron then decodes gene expression $\hat{X}_s$ influenced by these LR interactions, integrating knowledge from an established gene regulatory network. iMiracle isolates the baseline gene expression matrix $\hat{X}_b$ solely determined by cell types. The final imputed gene expression matrix $\hat{X}$ merges the baseline matrix with expressions from ligand-receptor interactions. Through iterative learning, $\hat{X}$ is used to progressively refine the multi-view graph, enhancing both imputation precision and the inference of ligand-to-gene regulatory scores.

pre-structured masks to boost biological relevance [9]. Second, iMiracle employs iterative learning to gradually fill in the missing edges in the cell-cell communication (CCC) network, circumventing the limitations posed by the sparsity of spatial transcriptomic data. Lastly, iMiracle infers a cell-specific ligand-gene regulatory score based on the contributions of different LR pairs to interpret inter-cellular regulation.

We applied iMiracle to ***nine*** simulated and ***eight*** real datasets across three sequencing technologies for comprehensive performance benchmarking. We found that iMiracle consistently outperforms ten methods in the gene expression imputation task and four methods in the regulatory score inference task. Lastly, we developed iMiracle into an open-source software package[1] to facilitate its use by the scientific community for investigating inter-cellular transcriptional regulation at the individual cellular level. With the rapid expansion of spatial transcriptomics data, we anticipate that iMiracle will be a powerful tool in decoding the nuances of CCC in complex tissues, thus enriching our understanding of inter-cellular-level ligand-gene regulatory impacts.

[1]https://github.com/aicb-ZhangLabs/iMiracle

2 Related Work

single-cell RNA sequencing (scRNA-seq) technology allows simultaneous gene expression profiling over thousands of cells, providing new opportunities to decipher inter-cellular transcriptional regulation [10, 21, 27, 42, 50]. Numerous methods have emerged to construct CCC networks based on ligand-receptor (LR) expression profiles [25]. While useful at their onset, they only focus on inter-cellular communication probabilities and do not delve into the transcriptional impacts on receiver cells. Later on, several methods were proposed to fill this gap by combining inter- and intra-cellular communications to link ligand genes from the sender cells directly to the target genes of the receiver cells. For example, NicheNet [3] combines inter-cellular CCC networks with prior knowledge of intra-cellular signaling and GRN to predict ligand-target gene regulatory scores. Cytotalk [60], on the other hand, combines cell-type-specific gene-co-expression networks with CCC networks to infer the regulatory potential of ligands on target genes. However, a challenge persists: scRNA-seq experiments require cell dissociation from their native tissue context, posing difficulties for accurate cell-specific inter-cellular regulatory relationship inference.

Current advancements in spatial transcriptomics have enabled spatially resolved gene expression profiling, enhancing our ability to explore transcription regulation within their native microenvironments [22, 32, 36]. Therefore, several computational methods were developed to utilize this new type of data. For instance, HoloNet [29] employed a multiview GNN to reconstruct gene expression and utilized an attention mechanism to calculate cell-type level ligand-gene regulatory score. However, the inherently sparse nature of spatial transcriptomics presents challenges in fully delineating the CCC network, resulting in an incomplete understanding of inter-cellular gene regulation [1]. Furthermore, it still lacks the granularity needed to explore ligand regulatory impacts at the level of individual cells.

3 Method

3.1 Method overview

As shown in **Fig. 1**, our iMiracle model contains two key modules: 1) an iterative GNN for accurate ***gene expression imputation*** of individual cells using a multi-view CCC network among LR pairs; 2) cell-specific ***regulatory score inference*** from ligand genes (in sender cells) to target genes (in receiver cells). Formally, given the observed sparse cell-by-gene matrix $\mathbf{X}_{obs} \in \mathbb{R}^{n\times m}$ (n cells and m genes), the spatial coordinates $\mathbf{C} \in \mathbb{R}^{n\times 2}$, and the cell type information $\mathbf{T} \in \mathbb{R}^{n\times t}$ (t is the number of cell types), iMiracle imputes the dense gene expression matrix as $\hat{\mathbf{X}} \in \mathbb{R}^{n\times m}$ and provides a ranked list $\mathbf{LR}_Y[c, g]$ to infer the regulatory score for cell c and gene g.

In its imputation module, iMiracle uniquely breaks down gene expression $\hat{\mathbf{X}}$ into two distinct components: firstly, a cell-type-specific baseline expression $\hat{\mathbf{X}}_b \in \mathbb{R}^{n\times m}$, which is determined by the cell type, and secondly, a cell-specific expression $\hat{\mathbf{X}}_s \in \mathbb{R}^{n\times m}$, which is influenced by the micro-environment through CCC. As shown in **Fig. 1**, iMiracle integrates a multi-view inter-cellular CCC network with either a Multi-Layer Perceptron (MLP) or an optional pre-defined GRN/PPI to predict $\hat{\mathbf{X}}$, $\hat{\mathbf{X}}_b$, and $\hat{\mathbf{X}}_s$. Then, iMiracle iteratively updates the LR pairs based on the imputed gene expressions, repeating the estimation process until convergence. In its second module, iMiracle infers ligand-target gene regulatory score based on the contribution of each LR pair to a gene of interest in a cell-specific manner. We will introduce the model details in the following sections.

3.2 Module 1: Gene expression imputation via an iterative GNN

iMiracle imputes the gene expression matrix $\hat{\mathbf{X}}$ without reference scRNA-seq data via three steps: constructing a multi-view CCC network, integrating inter- and intra-cellular networks, and iterative learning, as detailed below.

Multi-view CCC network construction. As shown in **Fig. 1**, for each LR pair, we calculate the communication probability for each cell by synthesizing gene expression information and spatial distance, represented by $\mathbf{G}_C^{lr}$. Then we combine all LR pairs' CCC information via a multi-view network $\mathbf{G}_C = \cup_{lr}\mathbf{G}_C^{lr}$, where $\cup$ is the view aggregation. The CCC construction requires three steps:

Step 1: identify expressed LR pairs. Starting with LR pairs from CellChatDB [25] (3,267 pairs for humans and 3,387 for mice), we extract the expression level for ligand l and receptor r from $\mathbf{X}_{obs}$. We define $\mathbf{S}_l$ and $\mathbf{S}_r$ as the sets of expression levels for l and r, respectively, and compute their geometric means as $\mathbf{E}_l = gmean(\mathbf{S}_l)$ and $\mathbf{E}_r = gmean(\mathbf{S}_r)$, each in $\mathbb{R}^{n\times 1}$. Then proportions of expressed cells are: $\xi_l = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\mathbf{E}_l[i] > 0\}$ and $\xi_r = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\mathbf{E}_r[i] > 0\}$. One LR pair is considered expressed if both ξ_l and ξ_r exceed the predefined threshold θ (set at 15% by default), forming the set of biologically active LR pairs:

$$\mathbf{LR}_\theta = \{lr | \xi_l > \theta \wedge \xi_r > \theta\}. \tag{1}$$

Step 2: calculate the distance for each cell. The Euclidean distance between cell $c1$ and cell $c2$ is calculated using their spatial coordinates ($\mathbf{C}[c1, :]$ and $\mathbf{C}[c2, :]$). This results in the distance matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$, capturing the spatial proximity of cells.

Step 3: compute the CCC network with gene expression. For each $lr \in \mathbf{LR}_\theta$, the CCC network $\mathbf{G}_C^{lr}$ is computed:

$$\mathbf{G}_C^{lr} = (\mathbf{E}_l \otimes \mathbf{E}_r) \odot \mathbf{D}^{-1}. \tag{2}$$

The outer product $\otimes$ yields a matrix where each entry signifies the combined expression of l and r for each cell pair. Element-wise multiplication $\odot$ merges this spatial data into the interaction strength assessment. Combining $\mathbf{G}_C^{lr}$ for all LR pairs results in the multi-view CCC network $\mathbf{G}_C$, enabling iMiracle to effectively model diverse CCC patterns.

Inter- and intra-cellular networks integration. iMiracle integrates inter- and intra-cellular networks to infer gene expressions in individual cells. For each lr pair, the GNN outputs node embeddings to capture CCC's impact from lr as

$$\mathbf{H}_{lr} = GNN(\mathbf{T}, \mathbf{G}_c^{lr}), \tag{3}$$

where $\mathbf{H}_{lr} \in \mathbb{R}^{n\times d}$ is the d-dimensional node embedding inferred from the lr-specific GNN. To model the intra-cellular regulation, iMiracle transforms $\mathbf{H}_{lr}$ into a gene expression matrix for each lr pair, followed by a shared decoder:

$$\hat{\mathbf{X}}_s^{lr} = Decoder(\mathbf{H}_{lr}). \tag{4}$$

Here $\hat{\mathbf{X}}_s^{lr} \in \mathbb{R}^{n\times m}$ reflects the gene expression regulated by the specific lr interactions. The decoder, typically implemented as an MLP, is designed to map each cell's embedding to its gene expression vector. Optionally, iMiracle can integrate the pre-structured GRN/PPI via:

$$\hat{\mathbf{X}}_s^{lr'} = \hat{\mathbf{X}}_s^{lr} \odot \mathbf{M}_{lr}, \tag{5}$$

where $\mathbf{M}_{lr} \in \mathbb{R}^m$ is a binary mask derived from the GRN/PPI, with ones representing possible regulation and zeros otherwise. In addition, iMiracle uses an MLP to capture baseline gene expression profiles that are solely influenced by cell type:

$$\hat{\mathbf{X}}_b = MLP(\mathbf{T}). \tag{6}$$

Then iMiracle synthesizes the cell-type-specific and cell-specific expression as the final gene expression $\hat{\mathbf{X}}$:

$$\hat{\mathbf{X}} = \hat{\mathbf{X}}_b + \sum_{lr\in\mathbf{LR}_\theta} \hat{\mathbf{X}}_s^{lr'}. \tag{7}$$

Iterative learning. Spatial transcriptomic data usually has excessive missing values in $\mathbf{X}_{obs}$, leading to incomplete CCC estimation and thus limiting the imputation performance. To address this issue, iMiracle employs iterative learning to gradually refine the

Table 1: Summary of simulation data parameters.

ID	k_b	r	(n_h, p_h)	(n_l, p_l)	(n_c, p_c)
1	2	10	(8, 0.5)	(2, 0.8)	(4, 0.8)
2	2	10	(8, 0.5)	(2, 0.8)	(8, 0.8)
3	5	10	(8, 0.5)	(2, 0.8)	(4, 0.8)
4	5	10	(8, 0.5)	(2, 0.8)	(8, 0.8)
5	5	10	(8, 0.5)	(4, 0.8)	(4, 0.8)
6	5	10	(8, 0.5)	(4, 0.8)	(8, 0.8)
7	5	20	(8, 0.5)	(2, 0.8)	(4, 0.8)
8	10	10	(8, 0.5)	(2, 0.8)	(8, 0.8)
9	20	10	(8, 0.5)	(2, 0.8)	(8, 0.8)

multi-view graph $\mathbf{G}_C$ based on the imputed expression matrix $\hat{\mathbf{X}}$. Specifically, after the i-th training iteration, $\mathbf{LR}_\theta^{(i+1)}$ is updated as:

$$\mathbf{LR}_\theta^{(i+1)} = \{lr | \xi_l^{(i+1)} > \theta \wedge \xi_r^{(i+1)} > \theta\}. \tag{8}$$

Here $\xi_l^{(i+1)}$ and $\xi_r^{(i+1)}$ represent the updated proportions of expressed cells, which are computed using the updated $\hat{\mathbf{X}}^{(i+1)}$. We next update the CCC network for each LR pair that exists in both $\mathbf{LR}_\theta^{(i)}$ and $\mathbf{LR}_\theta^{(i+1)}$. Combining previous CCC network $\mathbf{G}_c^{lr\,(i)}$, $\mathbf{G}_c^{lr\,(i+1)}$ is updated as:

$$\mathbf{G}_c^{lr\,(i+1)} = \alpha \mathbf{G}_c^{lr\,(i)} + (1-\alpha)\mathbf{G}_c^{lr\,(i+1)}. \tag{9}$$

A blending coefficient α harmonizes the contributions from both old and new estimates to ensure a smooth update. For LR pairs in $\mathbf{LR}_\theta^{(i+1)}$ but not in the $\mathbf{LR}_\theta^{(i)}$, they directly form new CCC networks: $\mathbf{G}_c^{lr\,(i+1)}$, which is derived from $\hat{\mathbf{X}}^{(i+1)}$. Merging existing and newly added CCC networks, we have:

$$\mathbf{G}_C^{(i+1)} = \cup_{lr \in \mathbf{LR}_\theta^{(i+1)}} \mathbf{G}_c^{lr\,(i+1)}. \tag{10}$$

Model training and hyperparameter tuning. During the training phase, iMiracle aims to minimize the Mean Squared Error (MSE) between $\hat{\mathbf{X}}$ and $\mathbf{X}_{obs}$. The loss function is particularly focused on non-zero entries of $\mathbf{X}_{obs}$:

$$\mathcal{L} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{1}[X_{obs,(i,j)} \neq 0](\hat{X}_{(i,j)} - X_{obs,(i,j)})^2}{\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{1}[X_{obs,(i,j)} \neq 0]}, \tag{11}$$

where $\mathbf{1}[\cdot]$ is an indicator that equals 1 if $X_{obs,(i,j)} \neq 0$ and 0 otherwise. We will stop the iteration if no new views can be added, as it suggests a saturation in constructing a full CCC.

We developed iMiracle using PyTorch version 1.12.1, operational on an Nvidia GeForce RTX A6000 GPU. Our computational setup is powered by an AMD EPYC 7302 16-Core Processor (1.0 TiB of memory) and operates on the Ubuntu 20.04.1 LTS system. In the gene imputation process, if there's no decrease for 10 consecutive epochs, we terminate the training and proceed to evaluate whether there's a need to update views. For the ligand-target gene regulatory score inference, training is halted if the loss fails to reduce by more than 0.001 over 10 successive epochs, after which we assess the necessity of updating views. For both tasks, we set the hidden dimension d to 32, the blending coefficient α to 0.2, the number of neighbors k to 5 for graph construction, a default two GNN layers, the maximum number of epochs to 1000, and use a learning rate of 0.01 with the Adam optimizer (details in the parameter analysis).

Table 2: Summary of real datasets.

Platform	Organism	Sample ID	Raw Matrix (Cell, Gene)	Raw Density	Filter Matrix (Cell, Gene)	Filter Density	# Imputed Entries
10xVisium	Human Dorsolateral Prefrontal Cortex (DLPFC)	151507	4226, 33538	0.042	4147, 4028	0.262	437240
		151508	4384, 33538	0.036	4148, 3342	0.258	358184
		151509	4789, 33538	0.043	4700, 4188	0.258	508186
		151510	4643, 33538	0.041	4547, 3908	0.259	461112
		151669	3661, 33538	0.054	3617, 5246	0.277	525930
		151670	3498, 33538	0.050	3433, 4909	0.272	457770
Stereoseq	Mouse	/	19109, 14376	0.024	4036, 1581	0.193	123444
SlideseqV2	Mouse	/	20139, 11750	0.031	5161, 2611	0.217	292418

3.3 Module 2: Cell-specific regulatory score inference

After training, iMiracle aims to identify ligands (in sender cells) that significantly impact gene expression (in receiver cells). For a specific cell c and gene g, the lr-related regulatory score $\psi(lr, c, g)$ is defined as:

$$\psi(lr, c, g) = \hat{\mathbf{X}}_s^{lr'}[c, g]. \tag{12}$$

Here $\hat{\mathbf{X}}_s^{lr'}[c, g]$ represents the lr-regulated strength for cell c and gene g. Based on $\psi(lr, c, g)$, iMiracle evaluates lr pairs within $\mathbf{LR}_\theta$ and gives a ranked list:

$$\mathbf{LR}_\gamma[c, g] = \text{sort}[lr \in \mathbf{LR}_\theta, \psi(lr, c, g) \text{ descending order}]. \tag{13}$$

$\mathbf{LR}_\gamma[c, g]$, ordered by regulatory score, enables iMiracle to pinpoint key LR pairs affecting gene regulation in individual cells, offering insights into inter-cellular regulation dynamics.

3.4 Simulation details

Following [29], we created simulated data, which includes 1000 cells in a 100-unit square space, for benchmarking. We assigned cell types based on their locations (using a parameter k_b that controls the mixing of cell types) and modeled gene expression for 50 genes (using a negative binomial distribution with high: (n_h, p_h) and low: (n_l, p_l)). To simulate CCC, 50 LR pairs were selected, with specific high-expression areas (a radius of r units and (n_c, p_c)) designated for intensified interactions. Gene expressions were updated to reflect these selected LR interactions. Next, we randomly masked the simulated data, maintaining a density of 20% to reflect the spatial data's sparsity. To ensure fairness, we designed nine different settings (**Table 1**) and reported performance across varied settings.

3.5 Data preprocessing and experimental setup

Preprocessing details. We include human dorsolateral prefrontal cortex (DLPFC) datasets from 10X Visium platform [35], mouse olfactory bulb dataset from Steroseq [5], and mouse olfactory bulb dataset from SlideseqV2 [48]. We follow pre-processing steps as suggested in the original paper (a summary can be seen in **Table 2**). Detailed methodologies for preprocessing and obtaining PPI and GRN are shown in the appendix.

Benchmark baselines and evaluation metrics. For the gene imputation task, data is down-sampled with 10% of non-zero entries allocated for testing and another 10% for validation [53]. To ensure fairness, this procedure is repeated ten times, each with different mask configurations. Imputed gene expressions are compared to ground truth using L1 Distance, Root-Mean-Square Error (RMSE), and Cosine Similarity. We evaluate ten leading methods,

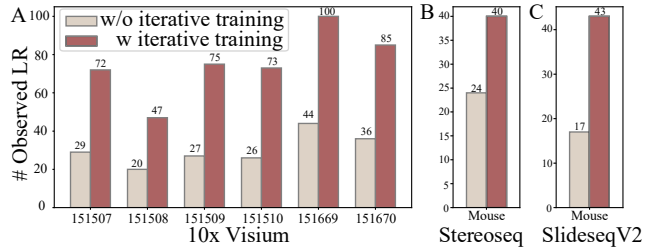

Figure 2: iMiracle fully delineates the CCC network via iterative learning, uncovering up to 181% more LR interactions in 10x Visium, 67% in Stereoseq, and 153% in SlideseqV2.

including scRNA-seq data analysis tools like scVI [34], ALRA [31], eSNN [49], MAGIC [51], and scGNN [53], which overlook spatial information. Additionally, gimVI [33] and Tangram [2], capable of integrating reference scRNA-seq, are tested in a reference-free mode for fairness. Spatial transcriptomics-specific methods like seSNN [43], STLearn [37], and STAGATE [8] are included.

For ligand-gene regulatory score inference using simulated data, we employed four evaluation metrics: Precision, Normalized Discounted Cumulative Gain (NDCG), Spearman Correlation, and Kendall Rank Correlation. Our comparison includes NicheNet [3], SpaTalk [44], and HoloNet [29], assessing their ability to rank LR pairs based on their influence on specific genes within cells, with a random guess approach as a naive baseline. We use default settings for all baseline methods.

4 Results

4.1 iMiracle delineates the full landscape of CCC network via iterative learning

To test the efficacy of iterative learning, we evaluated its role in the gradual delineation of the full landscape of CCC on eight datasets. Specifically, we compared the number of views in the constructed CCC, in other words, the number of included LR pairs. We found that iMiracle's iterative learning process noticeably increased the LR pairs included in G_C. For instance, on the 10x Visium datasets, iMiracle identified an increase of 27 to 56 LR pairs across six samples in the final iteration compared to the first round (**Fig. 2A**). This trend was consistent across all sequencing platforms, with an addition of 16 LR pairs in Stereoseq (**Fig. 2B**) and 26 in SlideseqV2 (**Fig. 2C**). The increased LR pair information enriched the spatial information in the GNN, potentially facilitating the downstream expression imputation and regulatory score inference tasks.

4.2 iMiracle consistently boosts imputation accuracy on diverse datasets

Next, we evaluated iMiracle's imputation performance against ten recent methods on diverse real datasets across three popular platforms (10x Visium, Stereoseq, and SlideseqV2) and two species (human and mouse). Due to the lack of gold standard benchmark datasets, we down-sampled the observed data and used the masked values as the ground truth to calculate three metrics, including L1 Distance, RMSE, and Cosine Similarity.

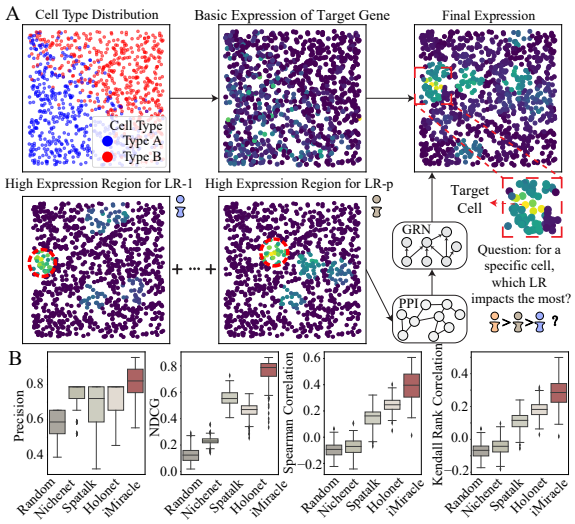

Figure 3: iMiracle consistently outperforms other methods in the regulatory score inference. (A) In this simulation, 1000 cells are spatially arranged in a 100-unit square. Cell types were determined by their locations, incorporating high-expression zones for various LR pairs, to realistically model gene expression and CCC dynamics. This setup is utilized for inferring cellular-level regulatory scores. (B) Benchmarking results demonstrate iMiracle's superior accuracy in inferring ligand-gene regulatory scores, surpassing all four baselines across all four metrics.

As shown in **Table 3**, iMiracle notably outperformed the best spatially-informed methods and demonstrated even larger improvements when compared to the top scRNA-seq-based baselines. On the SlideseqV2 dataset, for example, iMiracle achieved a 47% RMSE improvement over STAGATE, the foremost spatial method, and a 52% RMSE improvement over scGNN, the top non-spatial method. Specifically, among all methods utilizing cell coordinates, GNN-based approaches, such as STAGATE and iMiracle, demonstrated superior performance, supported by an average RMSE improvement of 66% over other spatial techniques. In addition, iMiracle exhibited higher imputation accuracy than STAGATE (RMSE 0.407 vs 0.765), attributable to its iterative learning and the multi-view network that combines both gene expression and distance information, as opposed to STAGATE's single-view GNN architecture derived mainly from the spatial distance. We also tested other datasets and found that iMiracle consistently reported the best gene imputation accuracy in all three metrics, indicating the robustness of our method across diverse sequencing platforms.

4.3 iMiracle highlights accurate inter-cellular ligand-gene regulatory insights

We benchmarked iMiracle with four other methods in terms of their ability to accurately capture ligand-target gene regulatory relationships across various simulated datasets (**Fig. 3A**, details see methods). Using known ligand-gene score as ground truth, we found that iMiracle consistently outperformed all the other methods

Table 3: Gene imputation benchmark. The best results are bolded. Results marked 'NA' for stLearn indicate unavailable HE stained images required by the method. "w/o" and "w" mean methods without and with spatial information, respectively.

Metric		Method	Platform & Dataset							
			10xVisium						**Stereoseq**	**SlideseqV2**
			DLPFC						Mouse	Mouse
			151507	151508	151509	151510	151669	151670	/	/
L1 Distance	w/o	scVI	0.794±0.004	0.838±0.006	0.800±0.002	0.670±0.003	0.810±0.003	0.696±0.005	1.442±0.005	1.127±0.006
		ALRA	0.499±0.003	0.512±0.001	0.490±0.001	0.496±0.001	0.467±0.002	0.472±0.002	0.406±0.013	0.649±0.066
		eSNN	1.254±0.001	1.373±0.001	1.266±0.001	1.294±0.000	1.017±0.001	1.071±0.001	2.802±0.002	2.071±0.002
		Magic	0.779±0.001	0.825±0.001	0.787±0.000	0.664±0.001	0.795±0.001	0.692±0.000	1.324±0.001	1.080±0.000
		scGNN	0.583±0.011	0.665±0.085	0.589±0.011	0.584±0.004	0.550±0.006	0.532±0.009	0.819±0.240	0.664±0.018
	w	gimVI	0.838±0.003	0.890±0.003	0.835±0.001	0.737±0.002	0.863±0.003	0.765±0.001	1.325±0.001	1.153±0.002
		seSNN	1.254±0.001	1.371±0.001	1.266±0.000	1.294±0.000	1.017±0.001	1.072±0.001	2.775±0.002	1.998±0.001
		Tangram	1.691±0.001	1.811±0.001	1.689±0.000	1.420±0.000	1.728±0.001	1.474±0.000	2.899±0.001	2.185±0.000
		STLearn	1.333±0.001	1.423±0.001	1.332±0.001	1.148±0.001	1.369±0.002	1.206±0.001.	NA	NA
		STAGATE	0.297±0.001	0.300±0.002	0.295±0.005	0.294±0.004	0.274±0.005	0.278±0.002	0.289±0.006	0.502±0.007
		iMiracle	**0.271±0.001**	**0.280±0.001**	**0.271±0.002**	**0.272±0.001**	**0.263±0.001**	**0.265±0.002**	**0.203±0.003**	**0.284±0.004**
Cosine Similarity	w/o	scVI	0.907±0.001	0.913±0.001	0.906±0.001	0.903±0.001	0.909±0.001	0.904±0.001	0.941±0.001	0.919±0.002
		ALRA	0.948±0.002	0.952±0.002	0.952±0.001	0.952±0.001	0.938±0.006	0.944±0.003	0.980±0.002	0.927±0.018
		eSNN	0.842±0.000	0.841±0.000	0.839±0.000	0.840±0.000	0.846±0.000	0.843±0.000	0.777±0.001	0.838±0.000
		Magic	0.915±0.000	0.920±0.000	0.914±0.000	0.909±0.000	0.916±0.000	0.910±0.000	0.968±0.002	0.936±0.000
		scGNN	0.933±0.004	0.927±0.016	0.932±0.002	0.936±0.000	0.917±0.002	0.929±0.002	0.948±0.035	0.953±0.002
	w	gimVI	0.957±0.000	0.965±0.001	0.955±0.001	0.947±0.001	0.962±0.001	0.948±0.002	0.964±0.000	0.936±0.001
		seSNN	0.843±0.000	0.841±0.000	0.840±0.000	0.841±0.000	0.851±0.000	0.847±0.000	0.768±0.000	0.817±0.000
		Tangram	0.713±0.001	0.725±0.001	0.717±0.001	0.716±0.001	0.717±0.001	0.715±0.000	0.772±0.001	0.763±0.001
		STLearn	0.718±0.000	0.718±0.000	0.715±0.001	0.724±0.000	0.715±0.001	0.717±0.000	NA	NA
		STAGATE	0.983±0.000	0.985±0.000	0.983±0.001	0.984±0.001	0.980±0.001	0.980±0.000	0.990±0.000	0.961±0.000
		iMiracle	**0.985±0.000**	**0.987±0.000**	**0.985±0.000**	**0.985±0.000**	**0.982±0.001**	**0.982±0.000**	**0.996±0.000**	**0.990±0.000**
RMSE	w/o	scVI	0.940±0.005	0.993±0.006	0.949±0.003	0.803±0.003	0.959±0.003	0.834±0.005	1.628±0.005	1.307±0.007
		ALRA	0.784±0.003	0.810±0.005	0.766±0.001	0.777±0.001	0.735±0.004	0.743±0.003	0.723±0.036	1.061±0.107
		eSNN	1.378±0.001	1.503±0.000	1.393±0.000	1.419±0.001	1.143±0.002	1.199±0.001	2.778±0.001	2.177±0.001
		Magic	0.917±0.001	0.972±0.001	0.929±0.000	0.792±0.000	0.936±0.001	0.824±0.000	1.453±0.001	1.238±0.001
		scGNN	0.755±0.016	0.850±0.096	0.762±0.011	0.755±0.002	0.717±0.007	0.686±0.010	1.051±0.307	0.842±0.021
	w	gimVI	0.955±0.002	1.002±0.001	0.957±0.001	0.858±0.001	0.970±0.002	0.890±0.002	1.448±0.001	1.217±0.004
		seSNN	1.354±0.001	1.474±0.000	1.370±0.000	1.395±0.001	1.119±0.001	1.175±0.001	2.770±0.002	2.087±0.001
		Tangram	1.768±0.001	1.889±0.001	1.767±0.000	1.503±0.000	1.804±0.001	1.557±0.000	2.970±0.001	2.284±0.000
		STLearn	1.516±0.001	1.629±0.001	1.521±0.001	1.300±0.001	1.556±0.002	1.362±0.001	NA	NA
		STAGATE	0.384±0.002	0.393±0.002	0.379±0.007	0.380±0.007	0.357±0.007	0.365±0.004	0.485±0.008	0.765±0.005
		iMiracle	**0.358±0.000**	**0.359±0.001**	**0.365±0.001**	**0.371±0.001**	**0.342±0.000**	**0.346±0.001**	**0.324±0.003**	**0.407±0.007**

(**Fig. 3B**). For instance, iMiracle demonstrated a noticeable improvement in NDCG (0.79 vs 0.24, **Fig. 3B**) when compared to NichNet, a gain largely due to its effective integration of spatial information. Among the spatial methods, iMiracle stood out as the best, surpassing SpaTalk and HoloNet (NDCG 0.79 vs 0.56/0.48, **Fig. 3B**). This trend was not only evident in NDCG but also consistent across other metrics such as precision, Spearman Correlations, and Kendall Rank Correlations. Such consistent performance highlights the benefit of using iterative learning to comprehensively map the CCC network, as well as its integration of both inter- and intra-cellular networks. This approach provides a more detailed, cell-specific view of cellular communication. Furthermore, iMiracle's improved performance was affirmed under various parameter settings (details in the appendix), underscoring the model's adaptability and effectiveness in diverse research contexts.

4.4 iMiracle reveals substantial regulatory heterogeneity across cells of the same type

One unique advantage of iMiracle is its ability to split gene expression into separate components driven by cell-type and the micro-environment, offering vital insights into how ligands differentially influence target genes within a specific spatial context. Therefore, our approach can quantitatively assess cell-specific spatial impacts of inter-cellular regulation and reveal regulatory variations among cells of the same type. We demonstrated this via a case study by estimating each LR's regulatory score to *GJA1*, a canonical marker gene in Astrocytes with essential functions in gap junction formation and DLPFC functionality [38, 47].

Specifically, we identified regions with high LR regulatory scores and gene expression of *GJA1* and intersected them with different layers, resulting in three unique regions to begin with (**Fig. 4A&E**). The top three LR pairs with the highest average regulatory scores were selected: *PTN-SDC4*, *APP-SORL1*, and *LRRC4B-PTPRD*. It's noteworthy that two out of three LR pairs (*PTN-SDC4* and *LRRC4B-PTPRD*) were identified via iterative learning, highlighting the importance of relying on dense imputed data.

We first compared the observed expression patterns with the two predicted components from iMiracle: cell-type-driven and micro-environment-driven expressions (**Fig. 4B-D**). These patterns showed high consistency in our visualizations. When we analyzed the cell-specific regulatory scores, we noticed substantial heterogeneity and distinct patterns for different LR pairs. For instance, the *PTN-SDC4* pair exhibited consistent scores across all regions, whereas the *APP-SORL1* and *LRRC4B-PTPRD* pairs showed strong preferences in specific regions (**Fig. 4F-H**). This finding underscores the importance of including cell-specific contexts in modeling processes, as relying solely on average cell-type-specific scores would obscure such significant regulatory diversity.

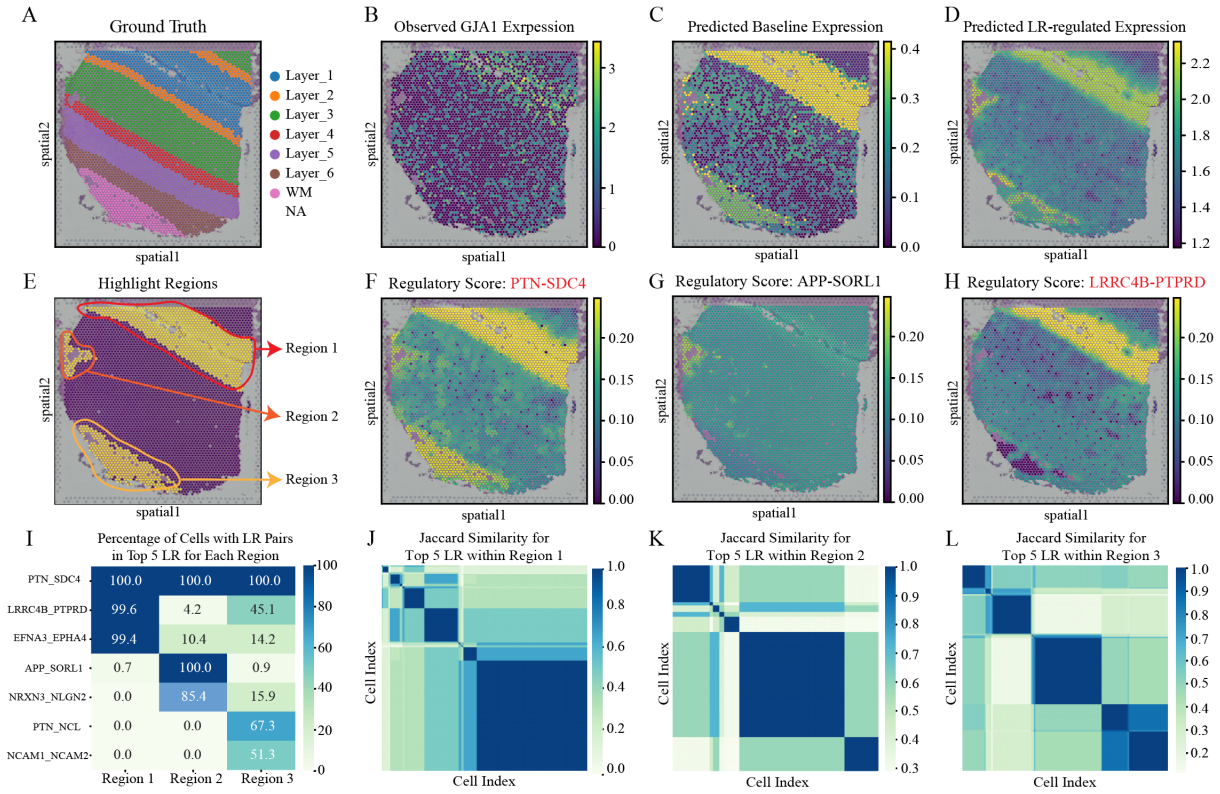

Figure 4: iMiracle reveals substantial regulatory heterogeneity across cells. (A) Detailed ground truth segmentation of the cortical layers and white matter (WM) within the DLPFC section of sample 151507. (B) Visualization of the observed expression pattern of *GJA1*. (C) Prediction of the baseline expression profile for *GJA1*. (D) Prediction of the LR-regulated expression for *GJA1*. (E) Identification of three key regions within sample 151507. (F-H) Top three LR interactions and their corresponding regulatory scores. LR pairs *PTN-SDC4* and *LRRC4B-PTPRD* were discovered through an iterative learning approach, indicated by their red colors. (I) Heatmap illustration of the percentage of cells featuring the top five LR pairs in each identified region. (J-L) *Jaccard similarity* of the top five LR pairs for cells within each region, revealing substantial regulatory heterogeneity across cells.

Finally, we compared both cross-region and within-region regulatory heterogeneity of the top 5 LR pairs. Only one LR pair *PTN-SDC4* was consistent across all three regions, while the remaining ones were highly regional-specific (**Fig. 4I**). For instance, *LRRC4B-PTPRD* pair ranked among the top 5 LR pairs in 99.6% of cells in region 1, whereas it was present in only 4.2% and 45.1% of cells in regions 2 and 3, respectively. Next, we looked at the regulatory heterogeneity within each region. Specifically, we calculated the *Jaccard similarity* of the identified top 5 LR pairs among cells within each region, as shown in **Fig. 4J-L**. Similarly, distinct LR usage preferences were discovered among cells within all three selected regions, demonstrating the pressing need to account for each cell's micro-environment when characterizing inter-cellular transcription regulation.

4.5 Ablation study to evaluate the effectiveness of iMiracle's modeling components

To assess each component of our model, we performed a variant analysis, considering four different versions: 1) "w/o GRN", which excludes the integration of prior biological knowledge; 2) "w/o iterations", a straightforward, non-iterative approach using sparse gene expressions; 3) "shared GNN", where the same GNN parameters are applied to all ligand-receptor (LR) pairs; and 4) "view decoder", implementing a unique decoder for each LR pair. This analysis allowed us to isolate and understand the individual contribution of each component to the overall performance of the model.

Firstly, after removing prior knowledge of intra-cellular network, we observed a slight decrease in gene imputation accuracy (1-11%, **Fig. 5A**) and a more pronounced reduction in regulatory score inference (NDCG: 0.43 vs 0.84, **Fig. 5B**). This outcome underscores the critical role of integrating biological knowledge for generating biologically meaningful interpretations. Next, the non-iterative model variant exhibited a slightly reduced accuracy in the regulatory score inference (NDCG: 0.78 vs 0.84, **Fig. 5B**), indicating the advantages of adopting an iterative approach. Then we found that employing shared GNN parameters led to a significant decline in gene imputation performance (25-32%, **Fig. 5A**), highlighting the

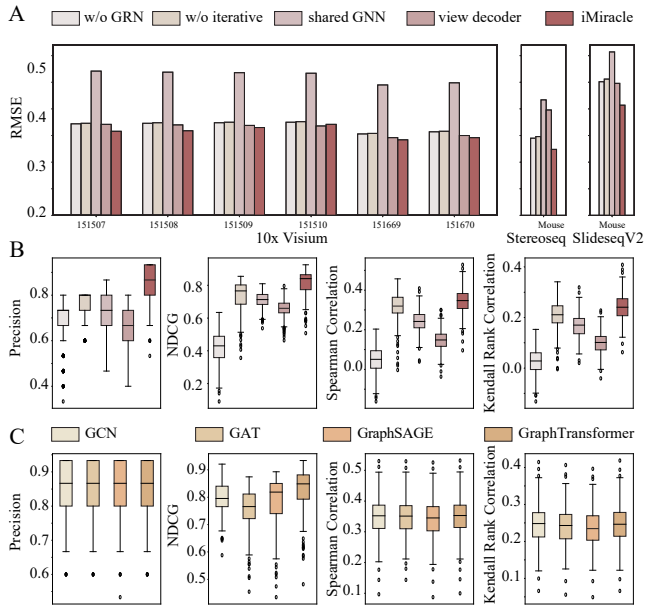

Figure 5: Variant analysis. (A) RMSE w.r.t. different variants of iMiracle for gene imputation. (B) Four regulatory score inference metrics w.r.t. variants of iMiracle. (C) Four regulatory score inference metrics w.r.t. GNN architectures.

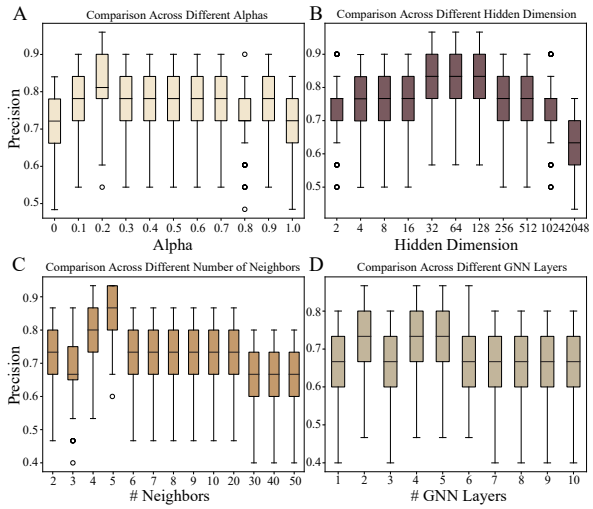

Figure 6: Parameter analysis. (A) Precision w.r.t. different blending coefficient α. (B) Precision w.r.t. different hidden dimension d. (B) Precision w.r.t. different number of neighbors k for graph construction. (B) Precision w.r.t. GNN layers L.

necessity for diverse message propagation strategies across different views in G_C. Lastly, using view-specific decoders adversely affected regulatory performance (NDCG: 0.66 vs 0.84, **Fig. 5B**), presumably due to the increased complexity in training arising from a higher number of parameters.

We also tested iMiracle's adaptability to different GNN architectures using GCN [26], GAT [52], GraphSAGE [20], and GraphTransformer [45]. Results showed comparable performance across these architectures (**Fig. 5C**), demonstrating iMiracle's flexibility and efficacy with various GNN models. We use GraphTransformer as our default setting.

4.6 Parameter analysis

To showcase the robustness of iMiracle in response to varying parameters, we employed the simulated data to assess its precision over a broad spectrum of blending coefficients α, hidden dimensions d, the number of neighbors k for graph construction, and the GNN layers L [12–14, 54, 55, 58]. As depicted in **Fig. 6A**, an α value of zero indicates exclusive reliance on newly imputed gene expressions for determining the existing graph structure. Conversely, an α value of one signifies maintaining the original graph structure of existing views. Both extremes lead to a reduction in precision. An α value of 0.2 results in optimal performance, underscoring the importance of smoothly integrating updated gene expression profiles into the multi-view graph. Exploring a wide range of hidden dimensions d, from 2 to 2048, we observed that iMiracle demonstrates considerable robustness in regulatory score inference, except at extreme values (i.e., 2, 1024, or 2048) from **Fig. 6B**. We choose 32 as the default d. Also, we set the number of neighbors $k = 5$ for graph construction, and the GNN layers $L = 2$ for optimal balance between performance and complexity as shown in **Fig. 6C-D**.

5 Conclusion and Discussion

In our study, we introduce iMiracle, a novel computational tool tailored for spatial transcriptomic data, aiming to unravel the complexities of inter-cellular transcriptional regulation. Unlike conventional methods that offer only averaged ligand regulatory scores across diverse micro-environments, iMiracle uniquely identifies the effects of gene expression caused by neighboring cells using CCC, separating these from effects due to the inherent characteristics of the cell type itself. This distinction enables iMiracle to investigate regulatory dynamics with unparalleled precision for a deeper understanding of inter-cellular transcriptional regulation.

iMiracle distinguishes itself from existing approaches via three key features designed explicitly for spatial transcriptomic data. Firstly, it integrates spatial distance and LR expression profiles to construct a multi-view inter-cellular CCC network, offering more biologically relevant insights with greater depth of information than methods mainly based on spatial distance (e.g., STAGATE). This integration, especially when combined with prior knowledge of intra-cellular networks (such as GRNs and PPIs), allows for more accurate and interpretable gene expression imputation, a benefit confirmed through extensive benchmarking on various datasets (**Table 3**). Secondly, iMiracle utilizes iterative learning to progressively refine the CCC network, effectively addressing data sparsity and uncovering more impactful LR pairs, as shown in our analyses on several real datasets (**Fig. 2**). Finally, it excels in inferring cell-specific ligand-gene regulatory scores, a feature often overlooked in approaches that neglect micro-environment effects (**Fig. 3**). We demonstrated the evident benefit of this feature by reporting substantial regulatory heterogeneity in cells under different spatial

contexts (**Fig. 4**). A minor concern regarding iMiracle is that it necessitates a cell-by-cell-type (or spot-by-cell-type-proportion) matrix as input to estimate baseline cell-type-specific gene expression. As a result, inaccuracies in cell type assignment or cell proportion calculation could affect the imputation performance. However, the impact of such inaccuracies is likely to be mitigated by ongoing and future advancements in spatial resolution and sequencing depth in technologies.

iMiracle has been developed as an open-source software freely available for researchers exploring inter-cellular gene expression regulation at the individual cellular level. Given the rapid advancements in spatial transcriptomics and the increasing availability of public data, iMiracle may serve as an essential tool in unraveling the complexities of cell-to-cell communication networks in complex tissues, thereby enriching our understanding of inter-cellular transcriptional regulation dynamics across various biological contexts.

ACKNOWLEDGEMENT

We would like to acknowledge support from the National Institutes of Health [R01HG012572, R01NS128523].

A Data Preprocessing Details

For real-world data preprocessing, we first filtered the cells and genes for quality assurance. Only cells with at least 500 detected genes and genes expressed in at least 10% of cells were retained. Next, we normalized the total counts per cell to a target sum of 1e4 and applied a log transformation to the data.

The protein-protein interaction (PPI) was obtained from a previous study named InWeb published in 2016 [30]. Specifically, we used the InWeb_InBioMap PPI table for Homo Sapiens which contained a total of 18,814 genes (vertices) and 883,356 interactions (edges). For the cell type-specific GRN in the brain prefrontal cortex region, we used a data processing pipeline from a set of private single-cell multiome data that contain scRNA-seq and scATAC-seq modalities. We first conducted a cellranger-arc call and basic QC to create the curated scRNA-seq matrix and the scATAC-seq fragment. Then, we used ArchR (v1.0.1) [19] to do cell type-specific peak calling and peak-to-gene interaction with functions addReproduciblePeakSet() and addPeak2GeneLinks(). In the peak calling step, we used Macs2 (version 2.2.7.1) [61] and did not limit the maximum number of peaks per cell type. In the peak-to-gene linkage creation step, we used the LSI created with 30 dimensions and used a correlation cutoff of 0.45 and a resolution of 500,000bp upstream and downstream. Then, we retrieve the motif-to-peak correspondence (motif annotation) from JASPAR2020 [18]. Using the annotation and the created peak-to-gene linkage, we construct the motif-to-gene graph if any peak connects to a gene and a motif.

To establish the mask from ligand to target gene, we processed genes associated with specific cell types and LR pairs. For each LR pair, genes were identified and expanded using PPI data to include associated genes. Subsequently, we integrated this information with GRN data to identify regulatory genes corresponding to each LR pair for different cell types. This process resulted in a comprehensive mapping, effectively linking LR pairs to their regulatory target genes, and thereby capturing the complex interplay within cellular communication networks.

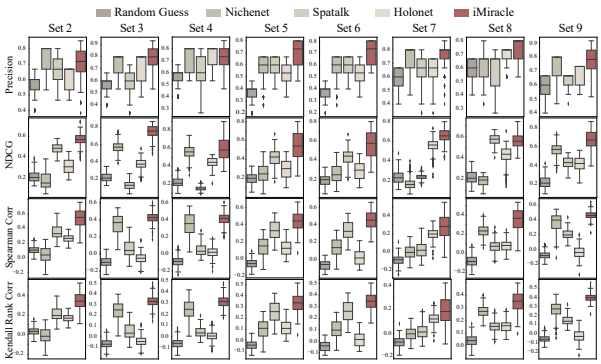

Figure 7: Additional regulatory score inference results.

B Complexity Analysis

We explored the complexity of iMiracle from two aspects: its time complexity, determined by various computational steps, and its parameter count, influenced by the distinct components.

B.1 Time Complexity Analysis of iMiracle

The time complexity of iMiracle is composed of four key steps: the basic gene expression decoupling, the multi-view graph construction, lr-specific GNNs, and the shared decoder. The basic MLP, processing cell type and gene expression data, incurs a complexity of $O(n \times (t \times d + d \times m))$. For the multi-view graph construction, each adjacency matrix $\mathbf{A}_{lr}$ computation entails a complexity of $O(n^2)$. With p unique LR pairs, the total complexity for this component is $O(p \times n^2)$. The computational load for each GNN layer is $O(|E| \times d)$, where $|E|$ denotes the number of edges in the sparse adjacency matrix. For all p LR pairs, this accumulates to $O(p \times |E| \times d)$. The decoders, applied to the embeddings from each lr-specific GNN, involve matrix operations resulting in a complexity of $O(p \times n \times (d \times d + d \times m))$. Considering $t \ll m$ and $d \ll m$, the overall time complexity of iMiracle can be summarized as $O(p \times n^2 + p \times d \times (|E| + n \times m))$.

B.2 The Number of Parameters in iMiracle

iMiracle's parameter complexity is influenced by its basic MLP, lr-specific GNNs, and decoders. The basic MLP comprises parameters of $O(t \times d + d \times m)$. For each lr-specific GNN, the parameter count is $O(t \times d + d \times d)$. With p unique LR pairs, the total parameters across all GNNs amount to $p \times O(t \times d + d \times d)$. Similarly, the shared decoder contributes an additional $O(d \times d + d \times m)$ parameters. Given that $t \ll m$ and $d \ll m$, the overall parameter count of iMiracle is $O(d \times (p \times (t + d) + m))$.

C Additional Performance Analysis

We benchmarked iMiracle with four other methods in terms of their ability to accurately capture ligand-target gene regulatory relationships under other eight simulation settings. As shown in **Fig. 7**, iMiracle consistently achieves the best performance across different settings and evaluation metrics.

References

[1] Jasim Kada Benotmane, Jan Kueckelhaus, Paulina Will, Junyi Zhang, Vidhya M Ravi, Kevin Joseph, Roman Sankowski, Jürgen Beck, Catalina Lee-Chang, Oliver Schnell, et al. 2023. High-sensitive spatially resolved T cell receptor sequencing with SPTCR-seq. *Nature communications* 14, 1 (2023), 7432.

[2] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Åsa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Ayshwarya Subramanian, Michal Lipinski, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Evan Z. Macosko, and Aviv Regev. 2021. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods* 18, 11 (01 Nov 2021), 1352–1362. https://doi.org/10.1038/s41592-021-01264-7

[3] Robin Browaeys, Wouter Saelens, and Yvan Saeys. 2020. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nature methods* 17, 2 (2020), 159–162.

[4] Yingxin Cao, Laiyi Fu, Jie Wu, Qinke Peng, Qing Nie, Jing Zhang, and Xiaohui Xie. 2021. SAILER: scalable and accurate invariant representation learning for single-cell ATAC-seq processing and integration. *Bioinformatics* 37, Supplement_1 (2021), i317–i326.

[5] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Jin Yang, Wenjiao Li, Jiangshan Xu, Shijie Hao, et al. 2021. Large field of view-spatially resolved transcriptomics at nanoscale resolution. *BioRxiv* 2021 (2021).

[6] Rujia Dai, Tianyao Chu, Ming Zhang, Xuan Wang, Alexandre Jourdon, Feinan Wu, Jessica Mariani, Flora M Vaccarino, Donghoon Lee, John F Fullard, et al. 2024. Evaluating performance and applications of sample-wise cell deconvolution methods on human brain transcriptomic data. *Science Advances* 10, 21 (2024), eadh2588.

[7] Chengyu Deng, Sean Whalen, Marilyn Steyert, Ryan Ziffra, Pawel F Przytycki, Fumitaka Inoue, Daniela A Pereira, Davide Capauto, Scott Norton, Flora M Vaccarino, et al. 2024. Massively parallel characterization of regulatory elements in the developing human cortex. *Science* 384, 6698 (2024), eadh0559.

[8] Kangning Dong and Shihua Zhang. 2022. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications* 13, 1 (2022), 1739.

[9] Ziheng Duan, Yi Dai, Ahyeon Hwang, Cheyu Lee, Kaichi Xie, Chutong Xiao, Min Xu, Matthew J Girgenti, and Jing Zhang. 2023. iHerd: an i ntegrative hie rarchical graph r epresentation learning framework to quantify network changes and prioritize risk genes in d isease. *PLoS Computational Biology* 19, 9 (2023), e1011444.

[10] Ziheng Duan, Cheyu Lee, Jing Zhang, et al. 2023. ExAD-GNN: explainable graph neural network for Alzheimer's disease state prediction from single-cell data. *APSIPA Transactions on Signal and Information Processing* 12, 5 (2023).

[11] Ziheng Duan, Dylan Riffle, Ren Li, Junhao Liu, Martin Renqiang Min, and Jing Zhang. 2024. Impeller: a path-based heterogeneous graph learning method for spatial transcriptomic data imputation. *Bioinformatics* 40, 6 (2024).

[12] Ziheng Duan, Yueyang Wang, Weihao Ye, Qilin Fan, and Xiuhua Li. 2022. Connecting latent relationships over heterogeneous attributed network for recommendation. *Applied Intelligence* 52, 14 (2022), 16214–16232.

[13] Ziheng Duan, Haoyan Xu, Yida Huang, Jie Feng, and Yueyang Wang. 2022. Multivariate time series forecasting with transfer entropy graph. *Tsinghua Science and Technology* 28, 1 (2022), 141–149.

[14] Ziheng Duan, Haoyan Xu, Yueyang Wang, Yida Huang, Anni Ren, Zhongbin Xu, Yizhou Sun, and Wei Wang. 2022. Multivariate time-series classification with hierarchical variational graph pooling. *Neural Networks* 154 (2022), 481–490.

[15] Ziheng Duan, Siwei Xu, Shushrruth Sai Srinivasan, Ahyeon Hwang, Che Yu Lee, Feng Yue, Mark Gerstein, Yu Luan, Matthew Girgenti, and Jing Zhang. 2024. scENCORE: leveraging single-cell epigenetic data to predict chromatin conformation using graph embedding. *Briefings in Bioinformatics* 25, 2 (2024), bbae096.

[16] Prashant S Emani, Jason J Liu, Declan Clarke, Matthew Jensen, Jonathan Warrell, Chirag Gupta, Ran Meng, Che Yu Lee, Siwei Xu, Cagatay Dursun, et al. 2024. Single-cell genomics and regulatory networks for 388 human brains. *Science* 384, 6698 (2024), eadi5199.

[17] Jonas Simon Fleck, Sophie Martina Johanna Jansen, Damian Wollny, Fides Zenk, Makiko Seimiya, Akanksha Jain, Ryoko Okamoto, Malgorzata Santel, Zhisong He, J Gray Camp, et al. 2023. Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* 621, 7978 (2023), 365–372.

[18] Oriol Fornes, Jaime A Castro-Mondragon, Aziz Khan, Robin van der Lee, Xi Zhang, Phillip A Richmond, Bhavi P Modi, Solenne Correard, Marius Gheorghe, Damir Baranašić, Walter Santana-Garcia, Ge Tan, Jeanne Chèneby, Benoit Ballester, François Parcy, Albin Sandelin, Boris Lenhard, Wyeth W Wasserman, and Anthony Mathelier. 2019. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* (Nov. 2019). https://doi.org/10.1093/nar/gkz1001

[19] Jeffrey M. Granja, M. Ryan Corces, Sarah E. Pierce, S. Tansu Bagdatli, Hani Choudhry, Howard Y. Chang, and William J. Greenleaf. 2021. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics* 53, 3 (Feb. 2021), 403–411. https://doi.org/10.1038/s41588-021-00790-6

[20] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).

[21] Lukas Heumos, Anna C Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D Lücken, Daniel C Strobl, Juan Henao, Fabiola Curion, et al. 2023. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics* (2023), 1–23.

[22] Miranda V Hunter, Reuben Moncada, Joshua M Weiss, Itai Yanai, and Richard M White. 2021. Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nature communications* 12, 1 (2021), 6278.

[23] Louise A Huuki-Myers, Abby Spangler, Nicholas J Eagles, Kelsey D Montgomery, Sang Ho Kwon, Boyi Guo, Melissa Grant-Peters, Heena R Divecha, Madhavi Tippani, Chaichontat Sriworarat, et al. 2024. A data-driven single-cell and spatial transcriptomic map of the human prefrontal cortex. *Science* 384, 6698 (2024), eadh1938.

[24] Shuaifei Ji, Mingchen Xiong, Huating Chen, Yiqiong Liu, Laixian Zhou, Yiyue Hong, Mengyang Wang, Chunming Wang, Xiaobing Fu, and Xiaoyan Sun. 2023. Cellular rejuvenation: molecular mechanisms and potential therapeutic interventions for diseases. *Signal Transduction and Targeted Therapy* 8, 1 (2023), 116.

[25] Suoqin Jin, Christian F Guerrero-Juarez, Lihua Zhang, Ivan Chang, Raul Ramos, Chen-Hsiang Kuan, Peggy Myung, Maksim V Plikus, and Qing Nie. 2021. Inference and analysis of cell-cell communication using CellChat. *Nature communications* 12, 1 (2021), 1088.

[26] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[27] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. 2020. Eleven grand challenges in single-cell data science. *Genome biology* 21, 1 (2020), 1–35.

[28] Brian D Lehmann, Antonio Colaprico, Tiago C Silva, Jianjiao Chen, Hanbing An, Yuguang Ban, Hanchen Huang, Lily Wang, Jamaal L James, Justin M Balko, et al. 2021. Multi-omics analysis identifies therapeutic vulnerabilities in triple-negative breast cancer subtypes. *Nature communications* 12, 1 (2021), 6276.

[29] Haochen Li, Tianxing Ma, Minsheng Hao, Wenbo Guo, Jin Gu, Xuegong Zhang, and Lei Wei. 2023. Decoding functional cell–cell communication events by multi-view graph learning on spatial transcriptomics. *Briefings in Bioinformatics* 24, 6 (2023), bbad359.

[30] Taibo Li, Rasmus Wernersson, Rasmus B Hansen, Heiko Horn, Johnathan Mercer, Greg Slodkowicz, Christopher T Workman, Olga Rigina, Kristoffer Rapacki, Hans H Stærfeldt, Søren Brunak, Thomas S Jensen, and Kasper Lage. 2016. A scored human protein–protein interaction network to catalyze genomic interpretation. *Nature Methods* 14, 1 (Nov. 2016), 61–64. https://doi.org/10.1038/nmeth.4083

[31] George C. Linderman, Jun Zhao, and Yuval Kluger. 2018. Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv* (2018). https://doi.org/10.1101/397588 arXiv:https://www.biorxiv.org/content/early/2018/08/22/397588.full.pdf

[32] Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, et al. 2023. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications* 14, 1 (2023), 1155.

[33] Romain Lopez, Achille Nazaret, Maxime Langevin, Jules Samaran, Jeffrey Regier, Michael I. Jordan, and Nir Yosef. 2019. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *CoRR* abs/1905.02269 (2019). arXiv:1905.02269 http://arxiv.org/abs/1905.02269

[34] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. 2018. Deep generative modeling for single-cell transcriptomics. *Nature Methods* 15, 12 (01 Dec 2018), 1053–1058. https://doi.org/10.1038/s41592-018-0229-2

[35] Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, Madhavi Tippani, et al. 2021. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience* 24, 3 (2021), 425–436.

[36] Reza Mirzazadeh, Zaneta Andrusivova, Ludvig Larsson, Phillip T Newton, Leire Alonso Galicia, Xesús M Abalo, Mahtab Avijgan, Linda Kvastad, Alexandre Denadai-Souza, Nathalie Stakenborg, et al. 2023. Spatially resolved transcriptomic profiling of degraded and challenging fresh frozen samples. *Nature Communications* 14, 1 (2023), 509.

[37] Duy Pham, Xiao Tan, Jun Xu, Laura F. Grice, Pui Yeng Lam, Arti Raghubar, Jana Vukovic, Marc J. Ruitenberg, and Quan Nguyen. 2020. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *bioRxiv* (2020). https://doi.org/10.1101/2020.05.31.125658 arXiv:https://www.biorxiv.org/content/early/2020/05/31/2020.05.31.125658.full.pdf

[38] Sovannarath Pong, Rakesh Karmacharya, Marianna Sofman, Jeffrey R Bishop, and Paulo Lizano. 2020. The role of brain microvascular endothelial cell and blood-brain barrier dysfunction in schizophrenia. *Complex Psychiatry* 6, 1-2 (2020), 30–46.

[39] Henry E Pratt, Gregory Andrews, Nicole Shedd, Nishigandha Phalke, Tongxin Li, Anusri Pampari, Matthew Jensen, Cindy Wen, PsychENCODE Consortium, Michael J Gandal, et al. 2024. Using a comprehensive atlas and predictive models to reveal the complexity and evolution of brain-active regulatory elements. *Science Advances* 10, 21 (2024), eadj4452.

[40] Jingyi Ren, Haowen Zhou, Hu Zeng, Connie Kangni Wang, Jiahao Huang, Xiaojie Qiu, Xin Sui, Qiang Li, Xunwei Wu, Zuwan Lin, et al. 2023. Spatiotemporally resolved transcriptomics reveals the subcellular RNA kinetic landscape. *Nature Methods* (2023), 1–11.

[41] W Brad Ruzicka, Shahin Mohammadi, John F Fullard, Jose Davila-Velderrain, Sivan Subburaju, Daniel Reed Tso, Makayla Hourihan, Shan Jiang, Hao-Chih Lee, Jaroslav Bendl, et al. 2024. Single-cell multi-cohort dissection of the schizophrenia transcriptome. *Science* 384, 6698 (2024), eadg5136.

[42] Fredrik Salmen, Joachim De Jonghe, Tomasz S Kaminski, Anna Alemany, Guillermo E Parada, Joe Verity-Legg, Ayaka Yanagida, Timo N Kohler, Nicholas Battich, Floris van den Brekel, et al. 2022. High-throughput total RNA sequencing in single cells using VASA-seq. *Nature Biotechnology* 40, 12 (2022), 1780–1793.

[43] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. 2015. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* 33 (2015), 495–502. https://doi.org/10.1038/nbt.3192

[44] Xin Shao, Chengyu Li, Haihong Yang, Xiaoyan Lu, Jie Liao, Jingyang Qian, Kai Wang, Junyun Cheng, Penghui Yang, Huajun Chen, et al. 2022. Knowledge-graph-based cell-cell communication inference for spatially resolved transcriptomic data with SpaTalk. *Nature Communications* 13, 1 (2022), 4429.

[45] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. 2020. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509* (2020).

[46] Disha Sood, Min Tang-Schomer, Dimitra Pouli, Craig Mizzoni, Nicole Raia, Albert Tai, Knarik Arkun, Julian Wu, Lauren D Black III, Bjorn Scheffler, et al. 2019. 3D extracellular matrix microenvironment in bioengineered tissue models of primary pediatric and adult brain tumors. *Nature communications* 10, 1 (2019), 4529.

[47] Joselyn S Soto, Yasaman Jami-Alahmadi, Jakelyn Chacon, Stefanie L Moye, Blanca Diaz-Castro, James A Wohlschlegel, and Baljit S Khakh. 2023. Astrocyte–neuron subproteomes and obsessive–compulsive disorder mechanisms. *Nature* (2023), 1–10.

[48] Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. 2021. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature biotechnology* 39, 3 (2021), 313–319.

[49] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. Comprehensive Integration of Single-Cell Data. *Cell* 177 (2019), 1888–1902. https://doi.org/10.1016/j.cell.2019.05.031

[50] Bram Van de Sande, Joon Sang Lee, Euphemia Mutasa-Gottgens, Bart Naughton, Wendi Bacon, Jonathan Manning, Yong Wang, Jack Pollard, Melissa Mendez, Jon Hill, et al. 2023. Applications of single-cell RNA sequencing in drug discovery and development. *Nature Reviews Drug Discovery* (2023), 1–25.

[51] David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, Kevin R. Moon, Christine L. Chaffer, Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. 2018. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* 174, 3 (2018), 716–729.e27. https://doi.org/10.1016/j.cell.2018.05.061

[52] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[53] Juexin Wang, Anjun Ma, Yuzhou Chang, Jianting Gong, Yuexu Jiang, Ren Qi, Cankun Wang, Hongjun Fu, Qin Ma, and Dong Xu. 2021. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nature communications* 12, 1 (2021), 1882.

[54] Yueyang Wang, Ziheng Duan, Yida Huang, Haoyan Xu, Jie Feng, and Anni Ren. 2022. MTHetGNN: A heterogeneous graph embedding framework for multivariate time series forecasting. *Pattern Recognition Letters* 153 (2022), 151–158.

[55] Yueyang Wang, Ziheng Duan, Binbing Liao, Fei Wu, and Yueting Zhuang. 2019. Heterogeneous attributed network embedding with graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 10061–10062.

[56] Cindy Wen, Michael Margolis, Rujia Dai, Pan Zhang, Pawel F Przytycki, Daniel D Vo, Arjun Bhattacharya, Nana Matoba, Miao Tang, Chuan Jiao, et al. 2024. Cross-ancestry atlas of gene, isoform, and splicing regulation in the developing human brain. *Science* 384, 6698 (2024), eadh0829.

[57] Yan Xia, Cuihua Xia, Yi Jiang, Yu Chen, Jiaqi Zhou, Rujia Dai, Cong Han, Zhongzheng Mao, PsychENCODE Consortium, Chunyu Liu, et al. 2024. Transcriptomic sex differences in postmortem brain samples from patients with psychiatric disorders. *Science Translational Medicine* 16, 749 (2024), eadh9974.

[58] Haoyan Xu, Ziheng Duan, Yueyang Wang, Jie Feng, Runjian Chen, Qianru Zhang, and Zhongbin Xu. 2021. Graph partitioning and graph neural network based hierarchical graph matching for graph similarity computation. *Neurocomputing* 439 (2021), 348–362.

[59] Lihua Zhang, Jing Zhang, and Qing Nie. 2022. DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. *Science Advances* 8, 22 (2022), eabl7393.

[60] Yang Zhang, Tianyuan Liu, Xuesong Hu, Mei Wang, Jing Wang, Bohao Zou, Puwen Tan, Tianyu Cui, Yiying Dou, Lin Ning, et al. 2021. CellCall: integrating paired ligand–receptor and transcription factor activities for cell–cell communication. *Nucleic acids research* 49, 15 (2021), 8520–8534.

[61] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* 9, 9 (Sept. 2008). https://doi.org/10.1186/gb-2008-9-9-r137