OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae716
https://doi.org/10.1093/bib/bbae716
Problem Solving Protocol

CellMsg: graph convolutional networks for ligand–receptor-mediated cell-cell communication analysis

Hong Xia[1], Boya Ji[1,*], Debin Qiao[2], Shaoliang Peng[1,*]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
[2]School of Computer and Artificial Intelligence, ZhengZhou University, Zhengzhou 450001, China

*Corresponding authors. Boya Ji, E-mail: byj@hnu.edu.cn; Shaoliang Peng, E-mail: slpeng@hnu.edu.cn

Abstract

The role of cell-cell communications (CCCs) is increasingly recognized as being important to differentiation, invasion, metastasis, and drug resistance in tumoral tissues. Developing CCC inference methods using traditional experimental methods are time-consuming, labor-intensive, cannot handle large amounts of data. To facilitate inference of CCCs, we proposed a computational framework, called CellMsg, which involves two primary steps: identifying ligand–receptor interactions (LRIs) and measuring the strength of LRIs-mediated CCCs. Specifically, CellMsg first identifies high-confident LRIs based on multimodal features of ligands and receptors and graph convolutional networks. Then, CellMsg measures the strength of intercellular communication by combining the identified LRIs and single-cell RNA-seq data using a three-point estimation method. Performance evaluation on four benchmark LRI datasets by five-fold cross validation demonstrated that CellMsg accurately captured the relationships between ligands and receptors, resulting in the identification of high-confident LRIs. Compared with other methods of identifying LRIs, CellMsg has better prediction performance and robustness. Furthermore, the LRIs identified by CellMsg were successfully validated through molecular docking. Finally, we examined the overlap of LRIs between CellMsg and five other classical CCC databases, as well as the intercellular crosstalk among seven cell types within a human melanoma tissue. In summary, CellMsg establishes a complete, reliable, and well-organized LRI database and an effective CCC strength evaluation method for each single-cell RNA-seq data. It provides a computational tool allowing researchers to decipher intercellular communications. CellMsg is freely available at https://github.com/pengsl-lab/CellMsg.

Keywords: cell–cell communications; ligand–receptor interactions; graph convolutional networks; three-point estimation method

Introduction

Cell–cell communications (CCCs) are vital for the development and maintenance of multicellular organisms and has crucial roles in numerous biological processes [1, 2]. For example, macrophages have long been recognized for their role in supporting erythroblast growth and development within the erythroid islands of the bone marrow [3]. Tumor-associated macrophages and cancer-associated fibroblasts (CAFs) play important roles in disease progression of the tumor microenvironment [4]. Therefore, a thorough analysis of CCCs during disease development and progression can enhance our understanding of disease mechanisms and aid in identifying new treatment strategies [5].

CCC is often facilitated by interactions between different proteins, such as receptor-receptor interactions, extracellular matrix-receptor interactions, and ligand–receptor interactions (LRIs) [6]. These interactions enable recipient cells to activate downstream signals through their corresponding receptors, leading to changes in transcription factor activity and gene expression [7, 8]. Specifically, the communication between two cells can be quantified by examining all LRIs involved in mediating this process [6].

The growing accessibility of scRNA-seq data and the valuable resources provided by LRI identifying studies have generated significant interest in the inference and analysis of CCCs [8–10]. However, experimental methods for studying CCCs are time-consuming, labor-intensive, and costly. Consequently, computational methods have become a valuable complement, allowing for the effective characterization of CCCs mediated by LRIs based on scRNA-seq data [11].

The process of analyzing and inferring CCC through computational methods primarily involves two primary steps: identifying LRIs and measuring the strength of LRIs-mediated CCCs [9, 12]. Therefore, it is of great value to establish a comprehensive, highly reliable, and well-organized LRI database [9]. On this basis, the classified LRI resources can be used to measure the strength of CCCs.

Recently, significant research efforts are directed towards constructing high-quality LRI databases. For instance, SingleCellSignalR [13] first computed the average expressions of the ligand and receptor, and then used their regularized product to calculate a score for each LRI. CellPhoneDB [14] employed a permutation

Received: September 9, 2024. **Revised:** December 4, 2024. **Accepted:** December 27, 2024

method to compute LRI scores based on ligand and receptor expression and used the resulting scores to evaluate LRI specificity. iTALK [15] detected significant LRIs by first identifying genes with high or differential expression, and subsequently matching and associating these genes using its LR database. In addition, CellDialog [5] took protein features as inputs, reduced the dimensionality of these features using a feature selection method based on tree-boosting and mixed effects models, and finally, LRIs were classified using KTBoost. Following the development of these LRI databases, numerous CCC inference methods have been developed using LRI databases. These methods can generally be categorized into four types: based on statistics, based on networks, based on spatial information, and based on tensors [5]. For example, CellPhoneDB [14] identified CCCs by detecting LRIs that are highly enriched between cell types. Network-based methods, such as Connectome [16], NATMI [17], and CellEnBoost [6], constructed complex network models to evaluate LRI-mediated CCCs. Connectome [16] provided two different edge weights, the first is the product of the normalized expression levels of the ligand and receptor in their respective cell types, the second is the average of the z-scores of their expression levels in these cell types. NATMI [17] used specific expression, expression product, and total expression to construct three distinct cell connectivity networks. CellEnBoost [6] utilized cell expression, specific expression, and expression product to infer CCCs. The methods based on spatial information, including stLearn [18], SpaOTsc [19], and Giotto [20], employed spatial data to elucidate the mechanisms of CCCs. The methods based on tensors, like scTensor [21] and Tensor-cell2cell [22], employed tensor decomposition techniques to analyze CCCs. LIANA [8] evaluated 7 methods and 16 CCC inference resources, along with the agreement between the predictions of these methods. Moreover, scHyper [23] predicted CCCs in single-cell RNA sequencing (scRNA-seq) data based on a hypergraph neural network. It learned static and dynamic embeddings to capture higher-order interaction patterns by constructing ligand–receptor pairs, sending cell types, and receiving cell types as hypergraph nodes. CPPLS-MLP [24] combined constrained partial least squares regression and multilayer perceptron to predict intercellular communication in single-cell and spatial transcriptomic data. By integrating gene expression, spatial coordinates and cell type label, the model calculated the strength and direction of intercellular communication.

Graph neural network (GNN) [25] is a deep learning model which can process the graph-structured data. Different from traditional models, GNNs can not only capture information of immediate neighbors for nodes but also perform deeper reasoning with multi-layers where each layer aggregates and refines hidden representations from previous layers as the learning proceeds. In this study, we designed a computational framework named CellMsg. CellMsg involves two primary steps: identifying LRIs and measuring the strength of LRIs-mediated CCCs. Specifically, CellMsg begins by extracting multimodal features of ligands and receptors to construct an initial feature matrix and constructing an adjacency matrix based on known associations of ligands and receptors. These matrices serve as inputs for a deep learning model, where features are extracted through two layers of GCNConv [26]. These features extracted by GCNs are then used as inputs for linear networks, leading to classification. To prevent over-smoothing and gradient vanishing issues during feature extraction in the GCN networks, we introduce skip connections. After LRI classification, CellMsg filters the identified LRIs (i.e. if the expression level of either the ligand or receptor in a pair is below a certain threshold or not expressed in a particular cell, the pair is considered not to mediate the corresponding CCC). The filtered LRIs are then used to calculate threshold result, product result, and cell result for each scRNA-seq data, and ultimately, CellMsg utilizes the three-point estimation method to measure CCC strength based on these three results. The flowchart of this framework is illustrated in Fig. 1. In summary, the contributions of this work are as follows: (i) a computational method called CellMsg proposed to analyze CCCs by identifying high-confident LRIs and using scRNA-seq data to measure the strength of CCCs that mediated by these LRIs. (ii) Compared with other methods of identifying LRIs, CellMsg uses graph convolutional networks (GCNs) for the first time and utilizes multimodal features of ligands and receptors as initial embeddings to obtain a more complete, reliable and well-organize LRI database for each scRNA seq data. (iii) CellMsg calculates the CCC strength using the three-point evaluation method through filtered LRIs, which obtains relatively accurate CCC analysis results in comparison with other popular tools and provides multiple visualizations. (iv) As a computational tool to facilitate researchers to analyze LRI-mediated CCCs, CellMsg provides public code examples, tutorials and documentation at github: https://github.com/pengsl-lab/CellMsg.

Materials and methods

Evaluation method

We evaluated the performance of CellMsg on LRI identification tasks using standard multiple categorization metrics, including accuracy (ACC.), precision (Prec.), recall (Rec.), F1 score (F1.), Matthews correlation coefficient (MCC), area under the precision-recall curve (AUPR), and area under the roc curve (AUC). The AUC is calculated by the area under the ROC curve, and the ROC curve is plotted by false positive rate (FPR) and true positive rate (TPR), while the AUPR is calculated by the area under the PR curve, and the PR curve can be plotted using the recall and precision. Given false positive (FP), true positive (TP), false negative (FN), true negative (TN), their formulas are as follows:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$TPR = recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1\_score = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

Additionally, we assessed the overlap of LRIs identified by CellMsg with the LRI databases of other CCC tools through the Jaccard index and LRI sensitivity. These two metrics were used to evaluate the reliability of LRIs identified by CellMsg and the definitions of the Jaccard index and LRI sensitivity are as follows:

$$Jaccard(set_i, set_j) = \frac{||(set_i \cap set_j)||}{||(set_i \cup set_j)||} \tag{7}$$

$$LRI_{sentivity}(set_i, set_j) = ||(set_i \cap set_j)|| \tag{8}$$

Where set_i and set_j represent sets of LRIs identified by two distinct CCC analysis method. $||\cdot||$ represents the number of elements in

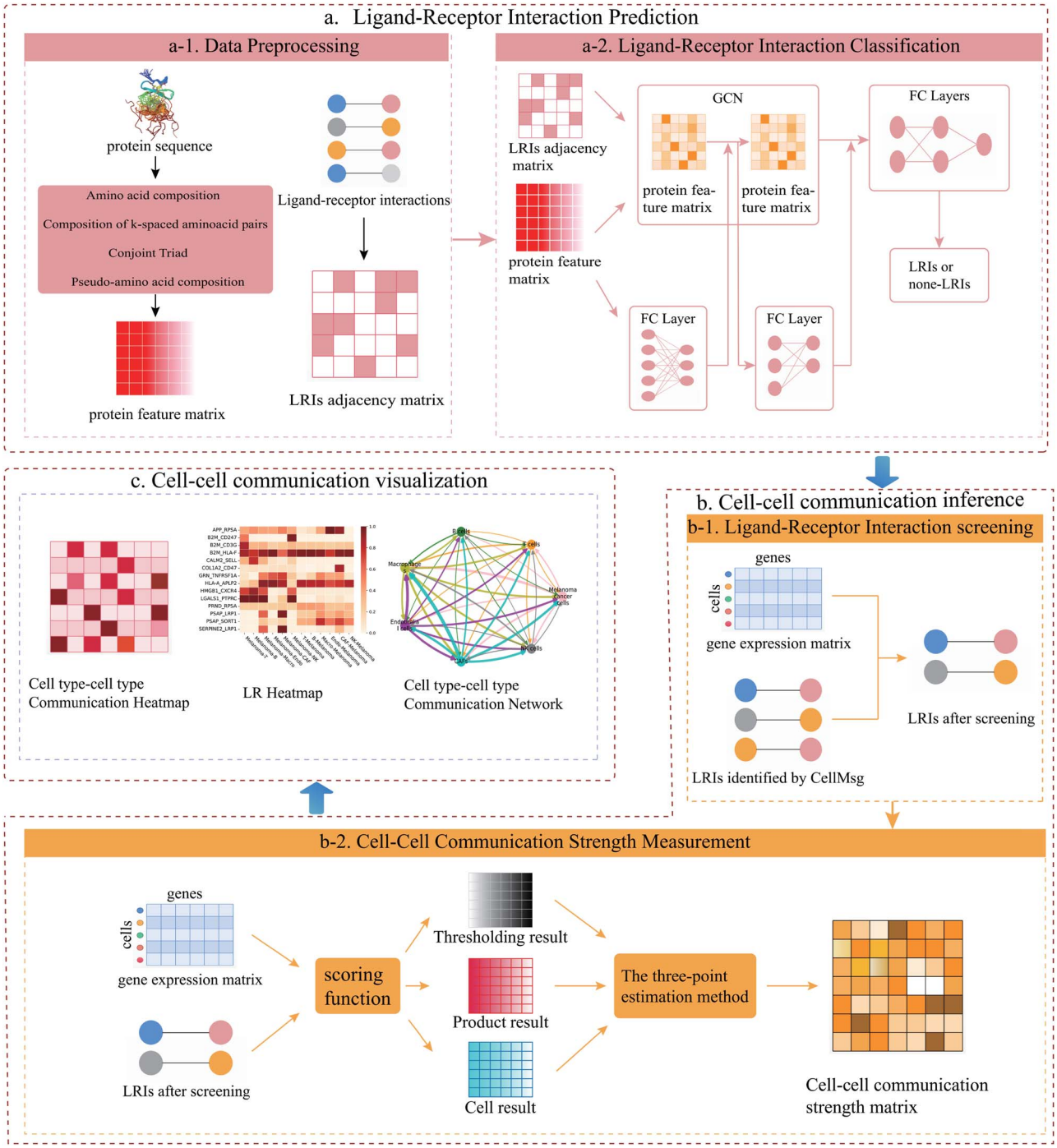

Figure 1. Workflow of the CellMsg method. (a) LRI prediction, including data preprocessing and LRI classification. The data preprocessing section extracts multimodal features of ligands and receptors to construct an initial feature matrix and constructs an adjacency matrix based on known associations of ligands and receptors. The LRI classification section uses these matrices as inputs to the GCNs to classify LRIs. (b) CCC inference, including LRI identification and CCC strength measurement. The LRI identification section filters high-confidence LRIs and then calculates the threshold result, product result, and cell result for each scRNA-seq data. On this basis, the CCC strength measurement section uses a three-point estimation method to calculate the CCC strength between different cell types. (c) CCC visualization, including the communication heatmap between different cell types, the communication network between different cell types, and the communication heatmap of the most active LR pairs between different cell types.

the set. $set_i \cap set_j$ represents the intersection of them and $set_i \cup set_j$ represents the union of them.

Data preprocessing

To evaluate the performance of CellMsg on LRI identification task, We utilized four distinct LRI datasets arranged by Peng *et al.* [5]. Datasets 1 and 2 were sourced from CellTalkDB [27], which contain 3398 LRIs in humans and 2033 LRIs in mice from the STRING database [28], respectively. Dataset 3 includes 2009 mouse LRIs that were organized by Skelly *et al.* [29]. Dataset 4 includes 6638 human LRIs that were compiled by Ximerakis *et al.* [30]. Finally, duplicate LRIs from the UniProt database [31] and LRIs without sequence information were removed. This process resulted in four LRI datasets, as shown in Table 1.

Table 1. The information of four LRI datasets

Dataset	Ligands	Receptors	LRIs number
Dataset 1 (human)	812	780	3390
Dataset 2 (mouse)	650	588	2031
Dataset 3 (mouse)	574	559	2006
Dataset 4 (human)	1129	1335	6585

To characterize each LRI, first, we used UniProt database [31] to download the ligands and receptors sequences. Next, the numerical features corresponding to the ligand and receptor sequences were extracted by iFeature [32]. These features include 2400 compositions of k-spaced amino acid pairs, 343 Conjoint Triad, 20 amino acid compositions, and 50 Pseudo-amino acid composition. As a result, a ligand or receptor is represented by a 2813-dimensional feature vector. Finally, we concatenated the feature matrices of ligands and receptors vertically.

In addition, we expanded the initial LRI matrix into a square matrix. Specifically, for dataset 1, we transformed the interaction matrix from dimensions 812×780 into a square matrix of dimensions 1592×1592. In this new matrix, the first 812 elements represent ligands, and the next 780 elements represent receptors. The values of the initial interaction matrix are placed in the upper right corner of the expanded matrix. Datasets 2, 3, and 4 are processed using the same steps.

LRIs prediction based on GCN

The GCN [26] is a deep learning model used for processing graph data. In GCN, the input consists of the adjacency matrix and the node feature matrix, which represent the graph structure. It has achieved success in numerous fields like bioinformatics, recommendation systems, and social network analysis. In CellMsg, we utilized GCNConv [26], a type of convolutional layer in GCNs. GCNConv determines the feature representation of each node based on its own features and the features of its neighboring nodes. Specifically, for each node, GCNConv aggregates the features of its neighbors and combines these aggregated features with the node's own features to generate an updated node feature representation. The computation process of each GCNConv layer can be represented as follows:

$$X' = D^{-\frac{1}{2}}\hat{A}D^{-\frac{1}{2}}X\theta + b \tag{9}$$

Here, D is a diagonal matrix, where D_{ii} represents the degree of the ith node in the adjacency matrix $\hat{A}$. Both θ and b are learnable parameters. X is the matrix obtained by concatenating the feature matrices of ligands and receptors. $\hat{A}$ represents the expanded LRI matrix with added self-loops, it can be represented as follows:

$$\hat{A} = A + E \tag{10}$$

where A is the expanded LRI matrix, and E is the identity matrix.

Additionally, we incorporated skip connections into each layer of GCNConv. This process involves applying a linear transformation to the input feature matrix, ensuring that the output feature matrix processed by GCNConv has the same dimensions as the original feature matrix. This allows the original and processed matrices to be added together, its formula is given as follows:

$$X^{l\_out} = XW + b' \tag{11}$$

where X is the feature matrix that is input to GCNConv. W and b' are both learnable parameters. Let m represent the feature dimension of the input feature matrix and n represent the feature dimension of the output feature matrix after a GCNConv layer, then the shape of W is $m \times n$, this transformation allows it to be added to the output feature matrix from GCNConv, addressing issues like over-smoothing or vanishing gradients when adding multiple layers of GCNConv to learn more comprehensive neighborhood information. Consequently, the formula for a GCNConv layer with skip connections is as follows:

$$X^{out} = ReLU((D^{-\frac{1}{2}}\hat{A}D^{-\frac{1}{2}}X\theta + b) + (XW + b')) \tag{12}$$

ReLU [33] is a widely used activation function in deep learning that introduces nonlinearity, enabling neural networks to learn and simulate complex function mappings. Its expression is expressed as follows:

$$f(x) = max(0, x) \tag{13}$$

Here, applying ReLU to a matrix means computing the maximum of 0 and each element in the matrix.

Finally, we employed three linear layers for the binary classification task of predicting LRIs. The first two linear layers, similar to the skip-connected layers, include a ReLU activation function after each layer. In the final linear layer, we set the output dimension to 1 and used a sigmoid [34] activation function to predict the probability of an interaction between ligands and receptors. By adjusting the threshold, we determined that 0.55 was the optimal value among 0.5, 0.55, and 0.6. Thus, if the predicted probability exceeds 0.55, we consider that there is an interaction between the ligand and receptor pair. The expression for the sigmoid function is as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{14}$$

Here, x is the output of the final linear layer.

Notably, we input the entire expanded LRI matrix and the protein feature matrix into GCNConv. Since the proportion of interacting ligands and receptors (positive samples with a value of 1 in the matrix) is very small compared to the entire matrix, we randomly selected an equal number of negative samples during model training after the final GCNConv layer. The feature vectors of the corresponding ligand and receptor in each sample were then concatenated to form the input to the linear layers.

Identifying high-confidence LRIs in scRNA-seq data

To measure the strength of CCC, we first used the trained model to predict all negative samples (i.e. where the value was 0 in the LRI matrix). If the interaction probability of a ligand–receptor (LR) pair exceeded a threshold β, it was considered a high-confidence LRI. In CellMsg, β was set to 0.999.

Next, we downloaded scRNA-seq data from the GEO database [35] for the queried tissue and further filtered the LRIs that were predicted high-confidenct and known by integrating them with scRNA-seq data. Specifically, if the ligand or receptor of an LRI was not expressed in the cells, we did not consider that the LRI mediated the corresponding CCC. This process yielded the LRIs that mediated CCCs in the queried tissue.

Computing threshold result, product result, and cell result through filtered LRIs

The expression threshold, cell expression, and expression product methods are three commonly used approaches for measuring the strength of CCC. The expression threshold method considers an LR pair 'active' only when the expression of the ligand in cell type C_1 and the expression of the receptor in cell type C_2 are both above a certain threshold. Specifically, for ligand l_a in cell type C_1 and receptor r_b in cell type C_2, we first calculated the mean expression levels M_a and M_b and standard deviations σ_a and σ_b of $l_a$and r_b across all cells. Then, we computed the mean expression level $M_{a,1}$ of l_a in cell type C_1 and the mean expression level $M_{b,2}$ of r_b in cell type C_2. If the mean expression level of l_a in C_1 is greater than $M_a + \sigma_a$, we considered l_a to be highly expressed in C_1. Similarly, if the mean expression level of r_b in C_2 is greater than $M_b + \sigma_b$, we considered r_b to be highly expressed in C_2. We then considered the LR pair to potentially mediate communication between the two cell types if both the ligand and the receptor are highly expressed in their respective cell types, as shown in the following formula:

$$LRI^{thr}_{a,1,b,2} = (M_{a,1} > M_a + \sigma_a) and (M_{b,2} > M_b + \sigma_b) \tag{15}$$

The expression product method infers CCC by directly multiplying the average expression value of ligand l_a in cell type C_1 with the average expression value of receptor r_b in cell type C_2. It calculates LRI-mediated CCC using the following formula:

$$LRI^{pro}_{a,1,b,2} = M_{a,1} \times M_{b,2} \tag{16}$$

The cell expression method infers CCC by multiplying the proportion of cells in cell type C_1, where the expression value of ligand l_a is greater than 0 by the proportion of cells in cell type C_2, where the expression value of receptor r_b is greater than 0. It is calculated for LRI-mediated CCC using the following formula:

$$LRI^{cell}_{a,1,b,2} = \frac{N_{a,1}}{N_1} \times \frac{N_{b,2}}{N_2} \tag{17}$$

Here, $N_{a,1}$ represents the number of cells in cell type C_1, where the expression value of ligand l_a is greater than 0, and N_1 represents the total number of cells in cell type C_1. Similarly, $N_{b,2}$ represents the number of cells in cell type C_2, where the expression value of receptor r_b is greater than 0, and N_2 represents the total number of cells in cell type C2.

Measuring CCC strength based on three-point estimation method

In the previous step, we obtained the CCC scores mediated by all filtered LRIs across different cell types using the expression threshold, cell expression and expression product methods. The final score of communication from cell type C_1 to cell type C_2 based on the expression threshold method was calculated by summing up the scores of all LRIs that mediate between cell type C_1 and cell type C_2, the expression is as follows:

$$score_{thr}(C_1, C_2) = \sum_{i=1}^{n} LRI^{thr}_{a,1,b,2} \tag{18}$$

where n represents the number of LRIs involved in mediating based on the expression threshold method, and a and b denote the ligand and receptor of an LRI, respectively.

For the cell expression and expression product methods, the final communication score from cell type C_1 to cell type C_2 was obtained similarly to the expression threshold method. We denote these scores as $score_{cell}(C_1, C_2)$ and $score_{pro}(C_1, C_2)$, respectively. In the scRNA-seq data we used, there are seven different cell types, which make up 49 possible cell type pairs. We calculated the final communication scores for these 49 pairs, denoted by $score_{thr}$, $score_{cell}$, and $score_{pro}$, representing the scores calculated by the expression threshold, cell expression and expression product methods, respectively. Next, we applied the min-max scaling method to normalize $score_{thr}$, $score_{cell}$ and $score_{pro}$, resulting in normalized CCC scores G_1, G_2, and G_3. We denoted the maximum, minimum, and median values in G_1, G_2, and G_3 as G_{max}, G_{min}, and G_{med} respectively. Finally, we computed the strength of CCC between any two cell types using the three-point estimation method [5]. Notably, the strength of CCC between two identical cell types was defined as 0. The expression for the three-point estimation method is as follows:

$$score = \frac{G_{max} + 4 \times G_{med} + G_{min}}{6} \tag{19}$$

Here, $score$ included the CCC strength between any two cell types obtained using the three-point estimation method.

CellMsg comprehensively considers threshold result, product result, and cell result. The threshold result filters out ligands or receptors whose average expression values in a cell type are lower than the average expression values in all cells, thus screening out important LRIs. The product result directly multiplies the average expression of ligands and receptors in the corresponding cell type, and since ligands and receptors tend to act synergistically in biology, the product method naturally mimics the synergistic effect. The cell result deeply consider the cellular specificity of the ligand or receptor by calculating the proportion of ligand or receptor that is non-negative expression in the corresponding cell type. Combining these three methods allows the strength of cellular communication to be quantified from a more integrated perspective.

Results and discussion

Evaluation of the LRI identification performance

In this section, we used five-fold cross-validation to divide the training set into five parts. Four parts were used for training, and the remaining part was used to validate the model. This process was repeated five times, with performance evaluated using the aforementioned metrics during each validation. Ultimately, we obtained five models. We calculated the average MCC m and the standard deviation b of the MCC for these five models. We then selected the models whose MCC fell within the range of $m \pm b$. Since the final goal is to use the trained model to predict potential LRIs, and some potential LRIs might exist in the initial LRI interaction matrix that have not yet been discovered, we selected the model with the highest recall from the chosen models for the final prediction task. As shown in the results in Table 2, we observed that the average values of all metrics under five-fold cross-validation across four datasets were robust. The average AUC across the four datasets reached 90%, and the average AUPR also approached 90%, demonstrating the robustness of CellMsg in predicting LRIs. Additionally, we plotted the ROC curves (Fig. 2) and PR curves (Fig. 3) for five-fold cross-validation on the four datasets to visually demonstrate the performance of CellMsg in the LRI identification task. The stability of the ROC and PR curves

Table 2. The average performance obtained by CellMsg on four datasets with five-fold cross-validation

Dataset	Acc.(%)	Prec.(%)	Rec.(%)	F1.(%)	MCC(%)	AUPR(%)	AUC(%)
dataset1	85.56	86.80	84.07	85.32	71.29	91.43	92.55
dataset2	83.56	83.60	83.51	83.48	67.22	89.33	91.09
dataset3	82.83	83.37	82.11	82.67	65.76	89.32	90.70
dataset4	83.31	83.60	82.96	83.22	66.71	90.24	91.42

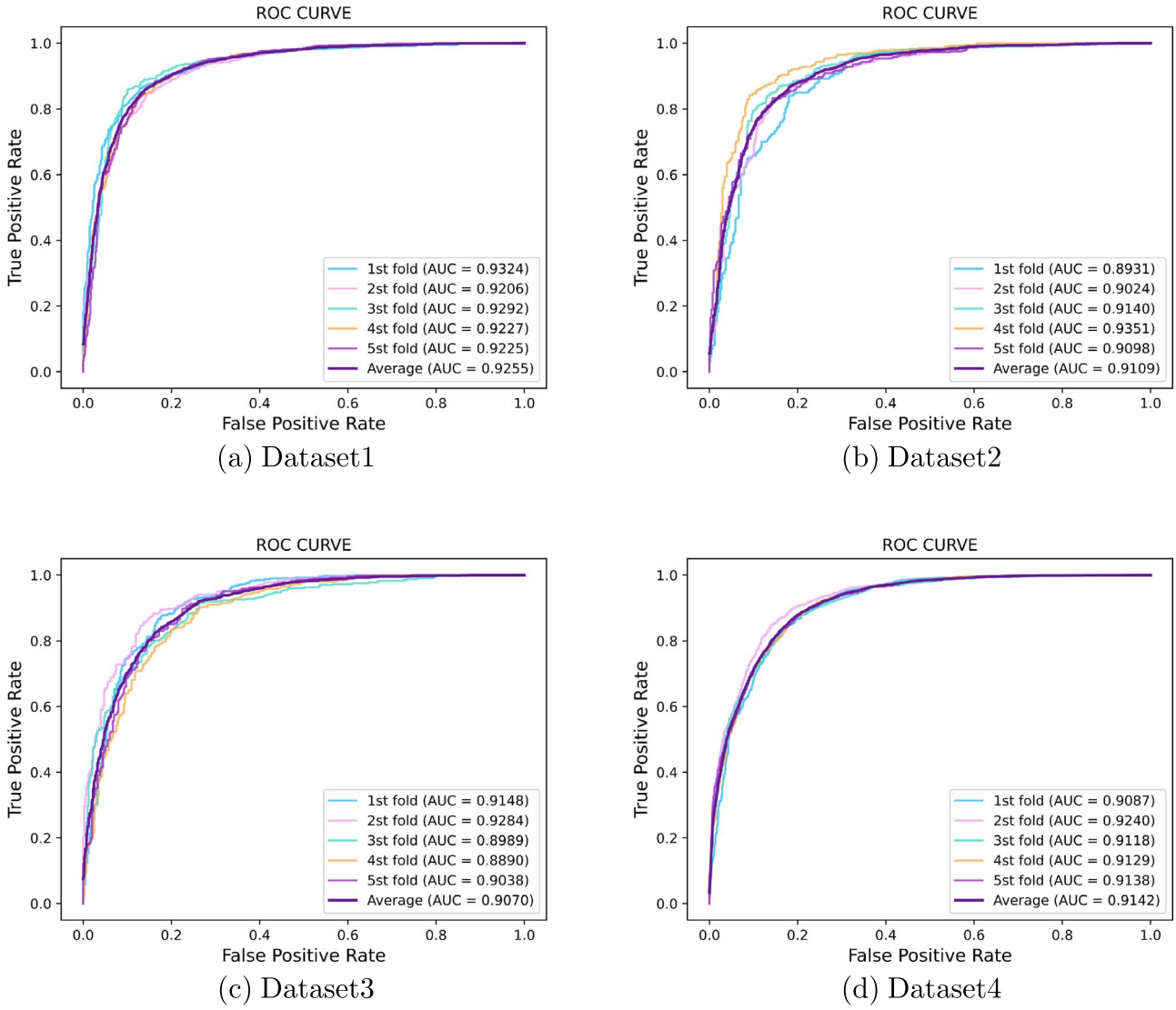

Figure 2. The ROC curves under five-fold cross validation obtained from CellMsg for LRIs prediction on Four LRI datasets. Datasets 1 and 4 provide human LRIs, Datasets 2 and 3 provide mouse LRIs.

in each validation further indicates the reliability of CellMsg in predicting LRIs.

Ablation study

Comparison between directed and undirected graphs

In the previously mentioned adjacency matrix A, the initial interaction matrix is located in the upper right corner of the expanded matrix, namely, we organized the adjacency matrix in the form of a directed graph. In this section, we reorganized it into an undirected adjacency matrix. Specifically, the initial interaction matrix resides in the upper right corner of the expanded matrix, and its transpose is located in the lower left corner.

We compared the LRI identification models under these two formats, and the results are shown in Table 3. It can be observed that the model based on the undirected adjacency matrix outperformed the directed one in terms of average performance across all datasets. Moreover, on datasets 2, 3, and 4, the model based on the undirected adjacency matrix exhibited lower standard deviations for the evaluation metrics, indicating greater stability. This superior performance is likely due to the undirected adjacency matrix's ability to facilitate more effective information aggregation in GCNs. By enabling symmetric information sharing, the undirected matrix captures reciprocal relationships between nodes (ligand-receptor pairs), leading to more robust feature

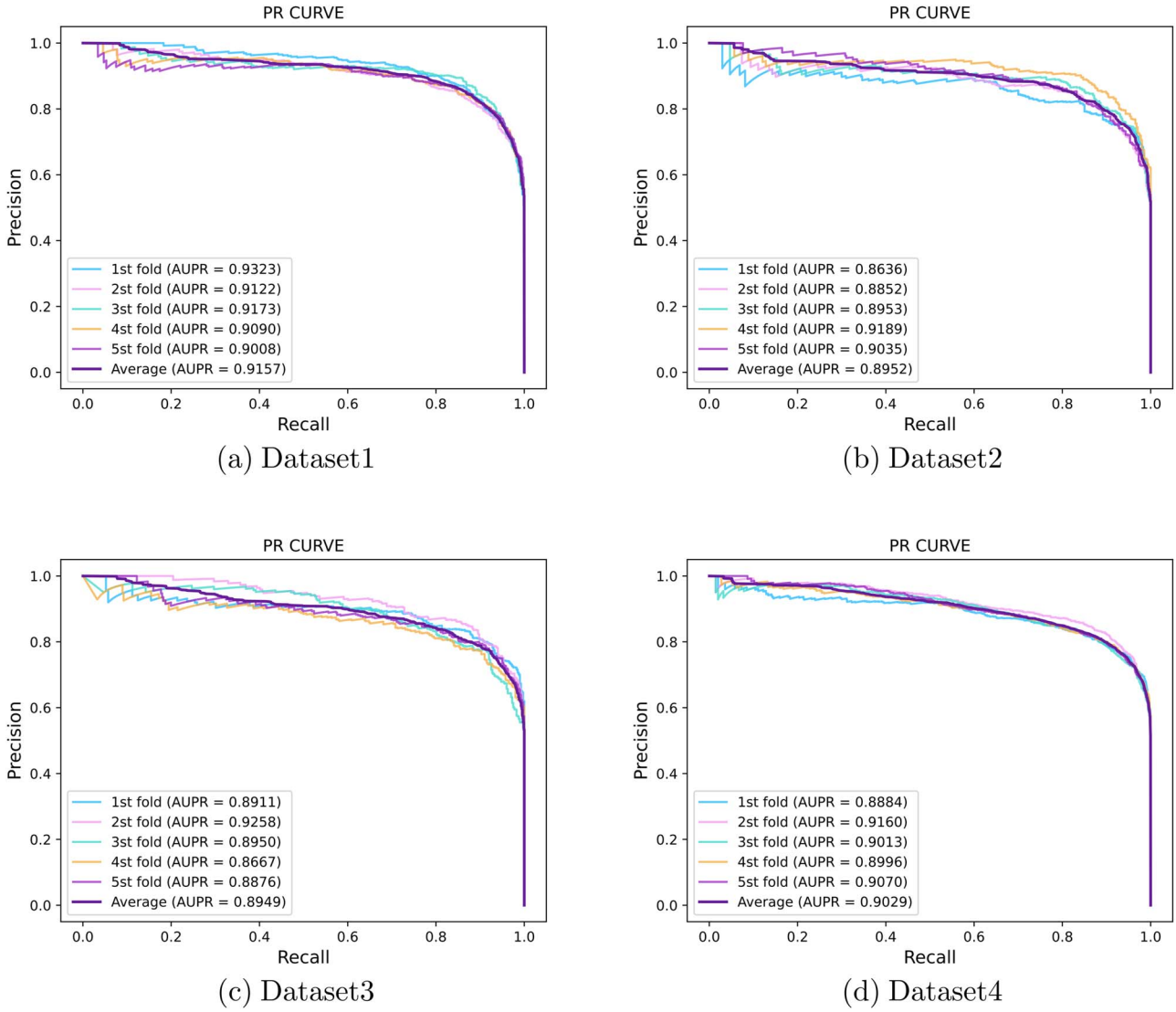

Figure 3. The PR curves under five-fold cross validation obtained from CellMsg for LRIs prediction on Four LRI datasets. Datasets 1 and 4 provide human LRIs, Datasets 2 and 3 provide mouse LRIs.

representations. Additionally, it allows for balanced and complete feature aggregation from neighboring nodes without directional constraints, improving the quality of learned features. Notably, our subsequent tasks were conducted using the model based on the directed adjacency matrix.

Comparison of the skip connections strategy

To further illustrate that the addition of skip connections to GCN can effectively improve the accuracy of the LRI prediction task, we introduced ablation experiments to verify the performance enhancement brought by skip connections. The results are shown in Table 4. From the results, it can be seen that for different datasets, CellMsg accompanied by the skip connection strategy shows better identification performance under different evaluation metrics.

Comparison with previous LRI Identification methods

We compared CellMsg with other leading LRI identification methods to further evaluate its performance, namely XGBoost [36], LightGBM [37], DNNXGB [38], CellEnBoost [6], and CellDialog [5]. XGBoost is a gradient boosting algorithm, while LightGBM implements a gradient boosting decision tree. DNNXGB [38] processes the features of ligands and receptors through two separate channels, and then concatenates the processed features, and inputs them into a fully connected layer for LRI identification. After training, the features output by the concatenation layer are further trained using XGBoost. CellEnboost [6] employs CNN and LightGBM for LRI identification and CellDialog [5] first uses GPBoost for feature selection and then uses KTBoost to identify LRI. We compared these methods on the same datasets, all using five-fold cross-validation. The average AUCs and AUPRs obtained by all methods are shown in Fig. 4. It is evident that our proposed CellMsg exhibits the best performance in the LRI identification task.

Validation of identified LRIs through molecular docking

In this part, we randomly selected a subset of LRIs identified by CellMsg in dataset 1 and performed rigid molecular docking of the ligands and receptors to validate whether these LR pairs are potential LRIs. We first obtained the 3D structure files of

Table 3. The LRI identification results under five-fold cross-validation on four datasets with CellMsg based on directed and undirected adjacency matrices

Metric	Dataset	Directed Adjacency Matrix	Undirected Adjacency Matrix
Acc.(%)	Dataset1	85.56 ± 1.10	87.54 ± 1.27
	Dataset2	83.56 ± 2.40	88.01 ± 0.38
	Dataset3	82.83 ± 1.53	87.01 ± 1.33
	Dataset4	83.31 ± 1.12	84.48 ± 0.90
AUC(%)	Dataset1	92.55 ± 0.45	94.83 ± 0.94
	Dataset2	91.09 ± 1.40	94.98 ± 0.71
	Dataset3	90.70 ± 1.36	94.20 ± 0.39
	Dataset4	91.42 ± 0.52	93.48 ± 0.50
AUPR(%)	Dataset1	91.43 ± 1.05	94.23 ± 1.00
	Dataset2	89.33 ± 1.85	93.99 ± 1.04
	Dataset3	89.32 ± 1.90	93.41 ± 0.58
	Dataset4	90.25 ± 0.91	92.92 ± 0.55
MCC(%)	Dataset1	71.29 ± 2.16	75.21 ± 2.48
	Dataset2	67.22 ± 4.71	76.11 ± 0.77
	Dataset3	65.77 ± 3.07	74.06 ± 2.66
	Dataset4	66.71 ± 2.19	69.33 ± 1.57

Table 4. Results of whether or not to introduce skip connections in GCNs

Metric	Dataset	With Skip Connections	Without Skip Connections
Acc.(%)	Dataset1	85.56	84.23
	Dataset2	83.56	82.37
	Dataset3	82.83	81.68
	Dataset4	83.31	81.73
AUC(%)	Dataset1	92.55	91.18
	Dataset2	91.09	90.01
	Dataset3	90.70	89.12
	Dataset4	91.42	90.05
AUPR(%)	Dataset1	91.43	89.17
	Dataset2	89.33	87.65
	Dataset3	89.32	87.15
	Dataset4	90.25	88.83
MCC(%)	Dataset1	71.29	68.67
	Dataset2	67.22	65.13
	Dataset3	65.77	63.78
	Dataset4	66.71	63.62

ligands and receptors from RCSB PDB [39], then converted the CIF files to PDB files through PDBj [40]. Next, we performed molecular docking using Gramm [41], and subsequently analyzed the generated PDB files of complexes using PDBePISA [42]. Lastly, we can obtain the interface area (IA) and binding energy (BE) of each LR pair. If the binding energy is less than -4 *kcal/mol*, it could be a potential LRI [5]. Table 5 presents the results of 10 randomly selected ligand-receptor pairs from our identified interactions. The binding energies of these 10 pairs are all less than -4 *kcal/mol*, indicating that they are likely potential LRIs.

Comparison of CellMsg with five CCC databases

In this part, we analyzed the overlap of LRIs between CellMsg and SingleCellSignalR [13], NATMI [17], CytoTalk [43], Connectome [16], and CellTalkDB [27] separately. We considered LR pairs identified by CellMsg with a probability greater than 0.999 as high-confidence LRIs (i.e. potentially existing LRIs) and merged these

Table 5. The molecular docking results of the LRIs we randomly selected

Ligand	Receptor	BE(kcal/mol)	IA(Å^2)
COL1A2	CD47	−23.7	1551.4
NTN1	ITGB1	−29.9	1750.9
ADAM10	ITGB1	−30.1	1755.4
TNC	ITGA1	−11	45.9
NLGN2	ITGB1	−29.4	1726.1
FBN1	ITGB8	−23.5	1832.8
FN1	ABCA1	−6.2	3105.6
COL4A2	ADGRB2	−31.3	2327.5
COL4A4	TNFRSF10A	−15.3	1035.1
CP	LRP2	−4.5	1900.0

with the known LRIs. As a result, CellMsg identified 4491, 3088, 2372, and 7837 LRIs on the four datasets, respectively. Table 6 presents the number of overlapping LRIs and the Jaccard index between CellMsg and the aforementioned CCC databases. It can be observed that the Jaccard indices between CellMsg and the other databases on Dataset 1 and Dataset 3 are almost all greater than or close to 50%, while the Jaccard index on Dataset 4 is less than 10%, indicating relatively fewer overlapping LRIs on Dataset 4. Additionally, since SingleCellSignalR and NATMI did not provide LRIs for mice, no comparison was made with these two tools on Dataset 2 and Dataset 3.

Comparison of CellMsg with existing CCC analysis methods

In this section, we conducted a comprehensive comparison between CellMsg and existing state-of-the-art CCC analysis methods, including CellPhoneDB [14], SingleCellSignalR [13], NATMI [17], Connectome [16], CellChat [44], and scHyper [23].

First, we performed the comparison based on a single-cell transcriptome data from human melanoma tissues derived from the GEO database [35] (accession code: GSE72056). Melanoma is a malignant tumor developed from the melanocytes of the epithelium of the skin and its appendages. Constructing CCC network at tumors initial stages is quite crucial for its diagnosis, prognosis and treatment. We investigated communications among seven cell types within the tissue: melanoma cancer cells, macrophages, CAFs, T cells, NK cells, endothelial cells, and B cells. Specifically, the human LRI database inferred by CellMsg was used for further filtering based on the single-cell expression profile. A LR pair is considered to mediate CCCs if the ligand and receptor are both expressed in the corresponding cell types. Based on the filtered LRIs, we calculated and visualized the CCC results inferred by CellMsg in melanoma tissues, as shown in Fig. 5. After that, we ranked the six cell types based on their communication results with melanoma cells as calculated from CellMsg and existing state-of-the-art CCC analysis methods, as shown in Table 7. We found that CellMsg inferred that CAFs have the strongest communication result with melanoma cancer cells in the microenvironment, which is the same as the results from CellPhoneDB [14], SingleCellSignalR [13], NATMI [17], Connectome [16], and scHyper [23]. In the tumor microenvironment, CAFs are one of the important cell types. They participate in tumor progression, metastasis, angiogenesis, reprogramming of immune cells, and resistance to therapy by providing extracellular matrix molecules, growth factors, cytokines, chemokines, and other regulatory molecules [45].

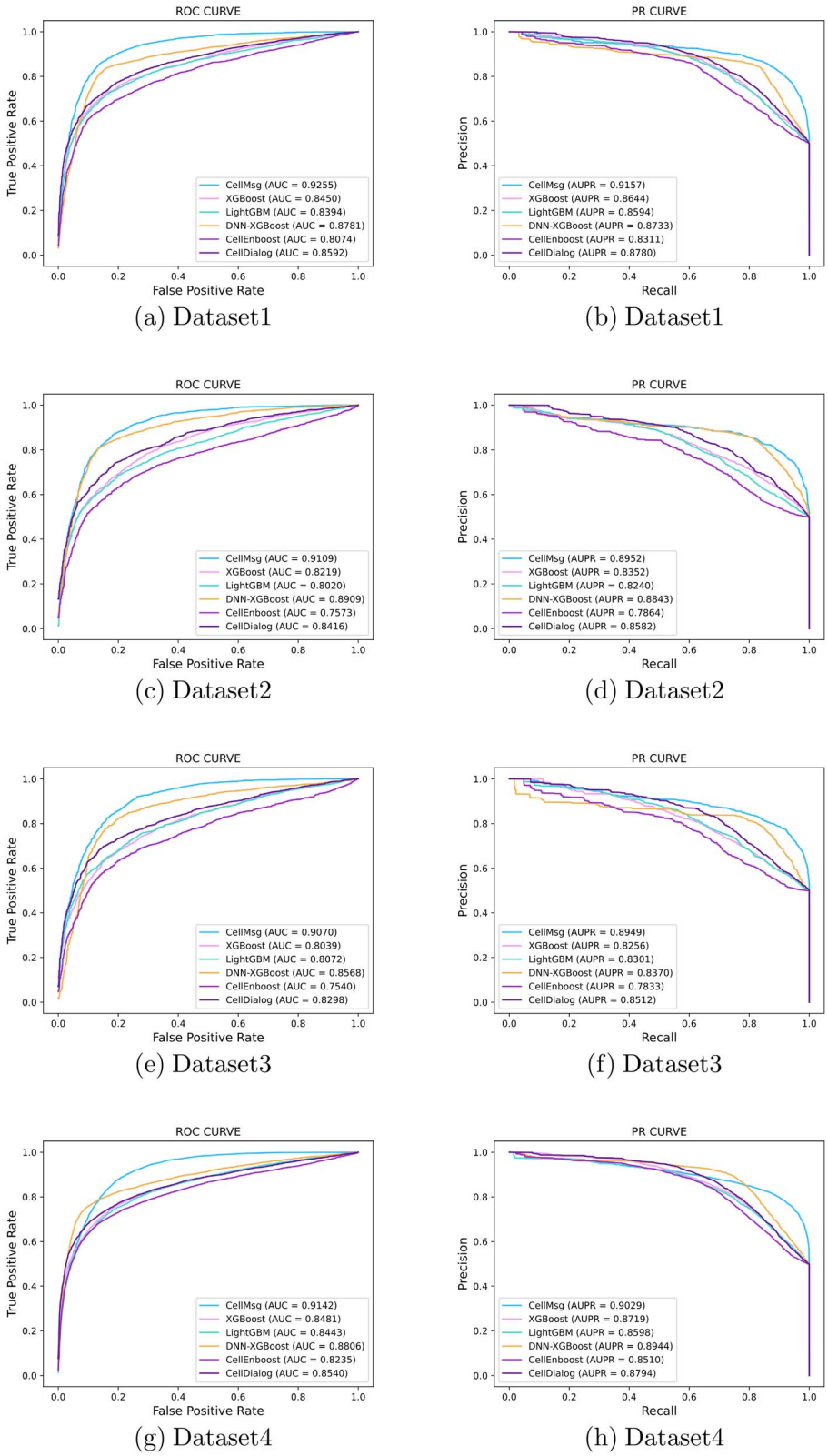

Figure 4. The ROC curves and PR curves obtained from CellMsg and the other five LRI identification models (XGBoost, LightGBM, DNN-XGBoost, CellEnBoost, and CellDialog) on four LRI datasets. Datasets 1 and 4 Provide Human LRIs, Datasets 2 and 3 Provide Mouse LRIs

Additionally, we analyzed the three most active ligand-receptor pairs in the communication between melanoma cancer cells and each of the six other cell types, as shown in Fig. 5(d). We organized these pairs into Table 8. It can be observed that the ligand-receptor pair B2M and HLA-F shows particularly active involvement in communication across multiple cell types, which aligns with fundamental immunological knowledge. B2M forms complexes with HLA molecules like HLA-F, playing crucial roles

Table 6. Comparison of LRIs identified by CellMsg with ones provided by five other CCC databases (the overlap number of LRIs/ the Jaccard index)

Dataset	SingleCellSignalR	NATMI	CytoTalk	Connectome	CellTalkDB	Total
Dataset1	2965/62.1%	1867/38.0%	1788/38.5%	2323/49.2%	3390/75.4%	3391/65.2%
Dataset2	/	/	1158/30.6%	1301/29.9%	2005/64.3%	2054/45.2%
Dataset3	/	/	1528/56.6%	1951/65.5%	1194/37.2%	1992/51.2%
Dataset4	483/4.8%	447/4.9%	440/5.0%	470/5.0%	501/4.9%	518/4.8%

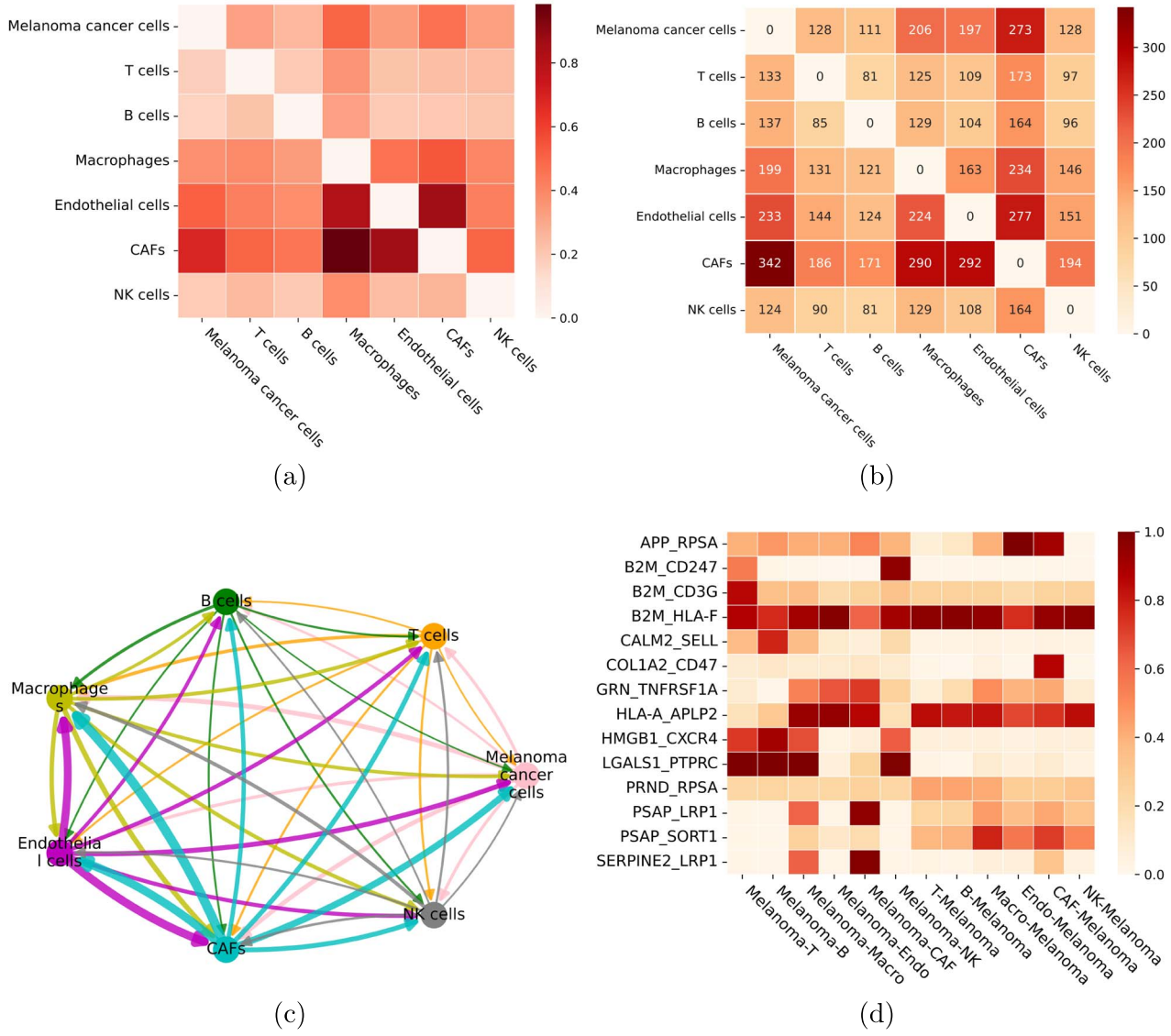

Figure 5. Visualization of the CCC results inferred by CellMsg in melanoma tissues. In (a) and (b), darker colors indicate stronger communication between the corresponding cell types; in (c), thicker edges represent stronger communication between cell types; in (d), the most active ligand-receptor pairs in the communication between melanoma cancer cells and six other cell types are shown, with darker colors indicating higher activity

in the immune system, particularly in antigen presentation processes [46].

For a more comprehensive comparison with other methods, we conducted a similar analysis based on a spatial transcriptome data from mouse kidney tissues derived from the STOmicsDB database [47] (Dataset ID: STDS0000121). We investigated communications among five cell types within the tissue: collecting duct principal cells (CDP), medullary cells, mesangial cells, progenitor cells and tubule cells (Fig. 6(a)). The medullary cells are distributed in the middle of the tissue. We calculated the communication results from medullary cells to other cell types by using CellMsg and existing state-of-the-art CCC analysis methods. From the spatial distribution of these cell types, we can easily find the spatial relationship of medullary cells with other cell types.

Table 7. Comparison of CellMsg with existing state-of-the-art CCC analysis methods in melanoma

Ranking	CellMsg	SingleCellSignalR	NATMI	CellPhoneDB	Connectome	CellChat	scHyper
1	CAFs	CAFS	CAFs	CAFs	CAFs	Macrophages	CAFs
2	Macrophages	Endothelial cells	Macrophages	Macrophages	Macrophages	CAFs	Endothelial cells
3	Endothelial cells	Macrophages	Endothelial cells	Endothelial cells	Endothelial cells	Endothelial cells	Macrophages
4	NK cells	T cells	NK cells	NK cells	NK cells	T cells	B cells
5	T cells	NK cells	T cells	T cells	T cells	NK cells	T cells
6	B cells	B cells	B cells	B cells	B cells	B cells	NK cells

Table 8. Top three active LRIs inferred by CellMsg ('in' denotes communication from other cell types to melanoma cancer cells, and 'out' denotes communication from melanoma cancer cells to other cell types, the content before and after the underscore represents the ligand and receptor of this LRI, respectively)

Cell type	'out' LRIs	'in' LRIs
T cells	LGALS1_PTPRC	B2M_HLA-F
	B2M_HLA-F	HLA-A_APLP2
	B2M_CD3G	PRND_RPSA
B cells	LGALS1_PTPRC	B2M_HLA-F
	HMGB1_CXCR4	HLA-A_APLP2
	CALM2_SELL	PRND_RPSA
Macrophages	LGALS1_PTPRC	B2M_HLA-F
	HLA-A_APLP2	HLA-A_APLP2
	B2M_HLA-F	PSAP_SORT1
Endothelial cells	B2M_HLA-F	APP_RPSA
	HLA-A_APLP2	B2M_HLA-F
	GRN_TNFRSF1A	HLA-A_APLP2
CAFs	SERPINE2_LRP1	B2M_HLA-F
	PSAP_LRP1	APP_RPSA
	HLA-A_APLP2	COL1A2_CD47
NK cells	LGALS1_PTPRC	B2M_HLA-F
	B2M_CD247	HLA-A_APLP2
	B2M_HLA-F	PSAP_SORT1

The mesangial, CDP, tubule and progenitor cells are sequentially distributed in the periphery of the medullary cells. Based on the assumption that spatially adjacent cell types should have stronger communication than spatially distant cells, we analyzed the results of different methods on the communication between other cell types and medullary cells (Fig. 6(b)–(h)). We found a clear association between the predicted communications and the spatial adjacency of their corresponding cell types for CellMsg, while the other methods showed inconsistent trends. Moreover, although both mesangial and CDP cells were distributed around medullary cells, CellMsg predicted stronger communication results for mesangial cells than for CDP cells, which was due to the higher number of mesangial cells captured by CellMsg. For tubule and progenitor cells, which have much higher cell numbers than mesangial and CDP cells, their communication strength with medullary cells is lower than that of mesangial and CDP cells due to their distance from medullary cells in their spatial location. Together, our analyses show that CellMsg performs well at predicting biologically meaningful communication in spatially adjacent cells than in distant cells from spatial transcriptomics datasets.

Conclusion

In this study, we present CellMsg method, which is an LRI-mediated CCC analysis method by incorporating LRI identification and filtering, CCC inference and visualization. Conceptually, CellMsg differs from these existing tools in terms of LRI identification and CCC inference, such as CellPhoneDB, SingleCellSignalR, NATMI, Connectome, CellChat, and so on. In terms of LRI identification, these existing tools make inference directly based on existing known LRI database, while CellMsg utilizes multimodal features of ligands and receptors and GCN to obtain a more complete, reliable, and well-organized LRI database. We demonstrate the accuracy of CellMsg in the identification of LRIs, demonstrate that CellMsg adopts the different GCN method with better performance in the identification of LRIs compared to these existing methods, and validate the accuracy of LRI identification by CellMsg using molecular docking method. In terms of CCC inference, we demonstrate that CellMsg obtains more accurate CCC analysis results than these existing tools based on single-cell transcriptome and spatial transcriptome data of different species and tissues. Our analyses show that CellMsg performs well at predicting biologically meaningful communication in spatially adjacent cells than in distant cells from spatial transcriptomics datasets. Practically, CellMsg is easy to implement and does not require complex operations and high computation resources. CellMsg needs no specific hardware resources, it can be used on a GPU but also in a CPU-only mode, which allows it to run on a broad range of hardware from desktop PCs to embedded systems. Once CellMsg has completed the inference of ligand-receptor pairs for a species, the scRNA-seq data of different tissues under the species does not need to be fine-tuned again. In the CCC analysis, we provide the completed ligand-receptor pairs for human and mouse inferred by CellMsg and the tutorial code on Github (https://github.com/pengsl-lab/CellMsg).

Furthermore, GCNs outperform other techniques in identifying LRIs, which may be attributed to the following features: (i) we calculate multimodal features of ligands and receptors as initial features of nodes in the GCN network, allowing the GCN to identify LRIs from multiple modalities. (ii) GCN captures direct and indirect relationships between ligands and receptors by combining graph topology information and attribute feature information through convolutional operations, thus being able to mine complex patterns of relationships from sparse interactions and effectively characterize and predict these interactions. (iii) The skip connection in the residual strategy is added to the GCN training process, which makes it possible to reduce the loss of information after each layer of convolution. It helps to improve

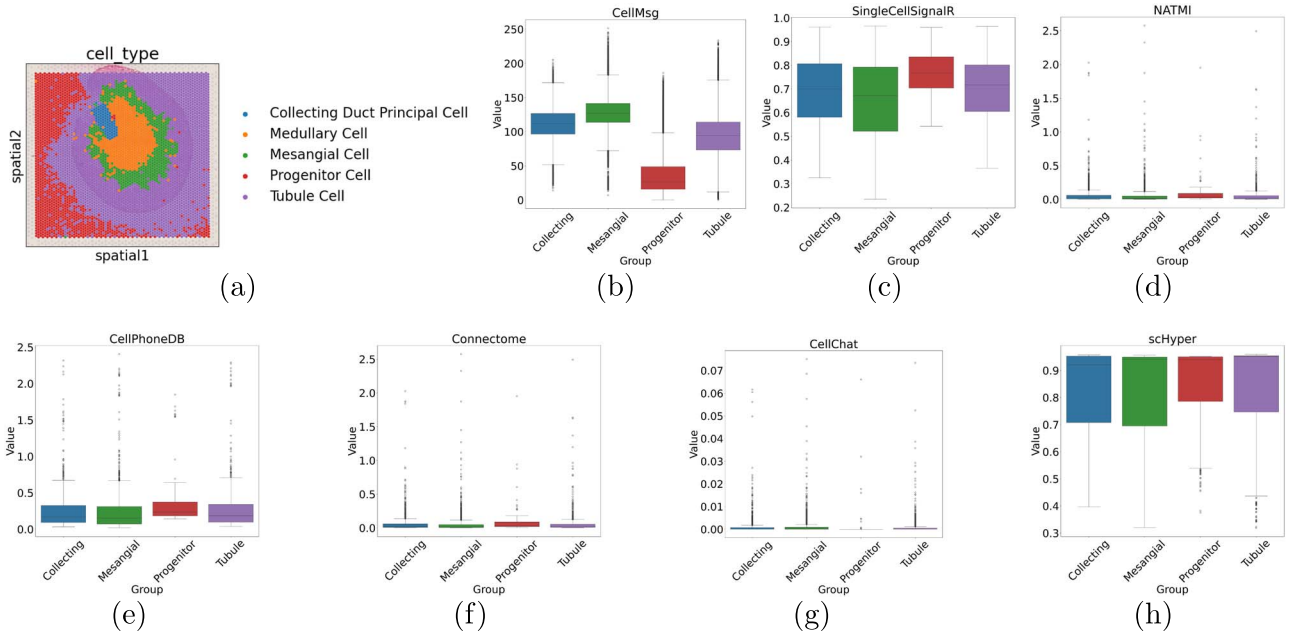

Figure 6. Comparison of the performance of CellMsg with existing state-of-the-art CCC analysis methods in mouse kidney tissue. (a) The UMAP of mouse kidney tissue shows the distribution of all cells. (b)–(h) Comparison of communication results between spatially adjacent and distant cells in the mouse kidney dataset.

the generalization ability of GCN and propagates the gradient more efficiently, making it easier to train and faster to converge.

As mentioned earlier, CellMsg can be efficiently run on both CPU and GPU, making it versatile for a wide range of hardware. While CellMsg works on standard hardware, GPU acceleration significantly enhances its performance, especially when dealing with large-scale single-cell datasets. In the future, CellMsg can be extended to common large scRNA-seq datasets with proper scaling techniques, such as data preprocessing (e.g. dimensionality reduction), subsampling and sparsifying, parallelization, memory-efficient operations, and cloud-based resources. These optimizations will ensure CellMsg can handle larger datasets efficiently, making it highly scalable for common scRNA-seq analyses.

In another area, CellMsg is theoretically applicable to scRNA-seq data from any species and its various tissues. While the tool relies on literature-supported ligand-receptor pairs for cellular communication, current ligand-receptor databases are primarily focused on human and mouse, such as the OminPath database. Therefore, our experiments were conducted using human and mouse datasets, but CellMsg can be extended to other species once comprehensive ligand-receptor data for those species become available. Importantly, once ligand-receptor pairs are inferred for a species, no further fine-tuning is required for different tissues within that species. However, CellMsg requires re-fine-tuning for different species and tissues during the CCC analysis phase. Detailed code and tutorials for both ligand-receptor inference and communication analysis are available on our GitHub.

Key Points

- The CellMsg method is a framework that analyze CCCs based on scRNA-seq data by identifying high-confident LRIs and measuring the strength of these LRIs-mediated CCCs.
- The CellMsg method introduces the graph convolutional network for the first time and uses multimodal features of ligands and receptors as initial embeddings to obtain a more complete, reliable, and well-organized LRI database for each single-cell RNA-seq data compared to other LRI identification methods.
- The CellMsg method calculates CCC strength by a three-point estimation method based on the filtered LRIs, which obtains relatively accurate CCC analysis results compared to popular CCC analysis tools and provides multiple visualizations.
- The CellMsg method serves as a computational tool to assist researchers in ligand-receptor-mediated CCC analysis. It provides public code examples, tutorials, and documentation available on github: https://github.com/pengsl-lab/CellMsg.

Funding

This work was supported by National Key R&D Program of China 2022YFC3400400, 2023YFC3503400; NSFC-FDCT Grants 62361166662; The Innovative Research Group Project of Hunan Province 2024JJ1002; Key R&D Program of Hunan Province 2023GK2004, 2023SK2059, 2023SK2060; Top 10 Technical Key Project in Hunan Province 2023GK1010; Key Technologies R&D Program of Guangdong Province (2023B1111030004 to F.F.H.). The Funds of State Key Laboratory of Chemo/Biosensing and Chemometrics, the National Supercomputing Center in Changsha, and Peng Cheng Lab; Graduate Research Innovation Project of Hunan Province (QL20230101).

References

1. Jonathan Singer S. Intercellular communication and cell-cell adhesion. *Science* 1992;**255**:1671–7. https://doi.org/10.1126/science.1313187.

2. Shao X, Xiaoyan L, Liao J. *et al.* New avenues for systematically inferring cell-cell communication: through single-cell transcriptomics data. *Protein&Cell* 2020;**11**:866–80. https://doi.org/10.1007/s13238-020-00727-5.
3. Roy CN, Mak HH, Akpan I. *et al.* Hepcidin antimicrobial peptide transgenic mice exhibit features of the anemia of inflammation. *Blood* 2007;**109**:4038–44. https://doi.org/10.1182/blood-2006-10-051755.
4. Kurashige M, Kohara M, Ohshima K. *et al.* Origin of cancer-associated fibroblasts and tumor-associated macrophages in humans after sex-mismatched bone marrow transplantation. *Commun Biol* 2018;**1**:131. https://doi.org/10.1038/s42003-018-0137-0.
5. Peng L, Xiong W, Han C. *et al.* Celldialog: A computational framework for ligand-receptor-mediated cell-cell communication analysis. *IEEE J Biomed Health Inform* 2023;**28**:580–91.
6. Peng L, Yuan R, Han C. *et al.* CellEnBoost: a boosting-based ligand-receptor interaction identification model for cell-to-cell communication inference. *IEEE Trans Nanobiosci* 2023;**22**:705–15. https://doi.org/10.1109/TNB.2023.3278685.
7. Peng L, Wang F, Wang Z. *et al.* Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies. *Brief Bioinform* 2022;**23**:bbac234. https://doi.org/10.1093/bib/bbac234.
8. Dimitrov D, Türei D, Garrido-Rodriguez M. *et al.* Comparison of methods and resources for cell-cell communication inference from single-cell RNA-seq data. *Nat Commun* 2022;**13**:1–13. https://doi.org/10.1038/s41467-022-30755-0.
9. Zhang Y, Liu T, Wang J. *et al.* Cellinker: a platform of ligand–receptor interactions for intercellular communication analysis. *Bioinformatics* 2021;**37**:2025–32. https://doi.org/10.1093/bioinformatics/btab036.
10. Xu J, Xu J, Meng Y. *et al.* Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data. *Cell Rep Methods* 2023;**3**:100382. https://doi.org/10.1016/j.crmeth.2022.100382.
11. Noël F, Massenet-Regad L, Carmi-Levy I. *et al.* Dissection of intercellular communication using the transcriptome-based framework icellnet. *Nat Commun* 2021;**12**:1089. https://doi.org/10.1038/s41467-021-21244-x.
12. Armingol E, Officer A, Harismendy O. *et al.* Deciphering cell–cell interactions and communication from gene expression. *Nat Rev Genet* 2021;**22**:71–88. https://doi.org/10.1038/s41576-020-00292-x.
13. Cabello-Aguilar S, Alame M, Kon-Sun-Tack F. *et al.* SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res* 2020;**48**:e55–5. https://doi.org/10.1093/nar/gkaa183.
14. Efremova M, Vento-Tormo M, Teichmann SA. *et al.* CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat Protoc* 2020;**15**:1484–506. https://doi.org/10.1038/s41596-020-0292-x.
15. Wang Y, Wang R, Zhang S. *et al.* iTALK: an R package to characterize and illustrate intercellular communication. BioRxiv2019;507871.
16. Raredon MSB, Yang J, Garritano J. *et al.* Computation and visualization of cell–cell signaling topologies in single-cell systems data using connectome. *Sci Rep* 2022;**12**:4187. https://doi.org/10.1038/s41598-022-07959-x.
17. Hou R, Denisenko E, Ong HT. *et al.* Predicting cell-to-cell communication networks using NATMI. *Nat Commun* 2020;**11**:5011. https://doi.org/10.1038/s41467-020-18873-z.
18. Pham D, Tan X, Xu J. *et al.* Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. *Nat Commun* 2023;**14**:7739. https://doi.org/10.1038/s41467-023-43120-6.
19. Cang Z, Nie Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat Commun* 2020;**11**:2084. https://doi.org/10.1038/s41467-020-15968-5.
20. Dries R, Zhu Q, Eng C-HL. *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* 2021;**22**:78. https://doi.org/10.1186/s13059-021-02286-2.
21. Tsuyuzaki K, Ishii M, Nikaido I. Uncovering hypergraphs of cell-cell interaction from single cell RNA-sequencing data. BioRxiv2019;**13**:566182.
22. Armingol E, Baghdassarian HM, Martino C. *et al.* Context-aware deconvolution of cell–cell communication with Tensor-cell2cell. *Nat Commun* 2022;**13**:3665. https://doi.org/10.1038/s41467-022-31369-2.
23. Li W, Wang H, Zhao J. *et al.* scHyper: reconstructing cell–cell communication through hypergraph neural networks. *Brief Bioinform* 2024;**25**:bbae436. https://doi.org/10.1093/bib/bbae436.
24. Zhang T, Zhenao W, Li L. *et al.* CPPLS-MLP: a method for constructing cell–cell communication networks and identifying related highly variable genes based on single-cell sequencing and spatial transcriptomics data. *Brief Bioinform* 2024;**25**:bbae198. https://doi.org/10.1093/bib/bbae198.
25. Scarselli F, Marco Gori A, Tsoi C. *et al.* The graph neural network model. *IEEE Trans Neural Netw* 2009;**20**:61–80. https://doi.org/10.1109/TNN.2008.2005605.
26. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: *ICLR 2017*, Toulon, France. 2017.
27. Shao X, Liao J, Li C. *et al.* CellTalkDB: a manually curated database of ligand–receptor interactions in humans and mice. *Brief Bioinform* 2021;**22**:bbaa269. https://doi.org/10.1093/bib/bbaa269.
28. Shao X, Liao J, Li C. *et al.* The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;**49**:D605–12.
29. Skelly DA, Squiers GT, McLellan MA. *et al.* Single-cell transcriptional profiling reveals cellular diversity and intercommunication in the mouse heart. *Cell Rep* 2018;**22**:600–10. https://doi.org/10.1016/j.celrep.2017.12.072.
30. Ximerakis M, Lipnick SL, Innes BT. *et al.* Single-cell transcriptomic profiling of the aging mouse brain. *Nat Neurosci* 2019;**22**:1696–708. https://doi.org/10.1038/s41593-019-0491-3.
31. U. Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**:D506–15. https://doi.org/10.1093/nar/gky1049.
32. Chen Z, Zhao P, Li F. *et al.* iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018;**34**:2499–502. https://doi.org/10.1093/bioinformatics/bty140.
33. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. *Proc Mach Learn Res* 2011;**15**:315–23.
34. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986;**323**:533–6. https://doi.org/10.1038/323533a0.
35. Barrett T, Suzek TO, Troup DB. *et al.* NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res* 2005;**33**:D562–6.
36. Wang X, Zhang Y, Yu B. *et al.* Prediction of protein-protein interaction sites through extreme gradient boosting with kernel principal component analysis. *Comput Biol Med* 2021;**134**:104516. https://doi.org/10.1016/j.compbiomed.2021.104516.

37. Chen C, Zhang Q, Ma Q. *et al.* Predicting protein-protein interactions through lightgbm with multi-information fusion. *Chemom Intel Lab Syst* 2019;**191**:54–64. https://doi.org/10.1016/j.chemolab.2019.06.003.
38. Mahapatra S, Gupta VRR, Sahu SS. *et al.* Deep neural network and extreme gradient boosting based hybrid classifier for improved prediction of protein-protein interaction. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**19**:155–65. https://doi.org/10.1109/TCBB.2021.3061300.
39. Ormo M, Cubitt AB, Kallio K. *et al.* Crystal structure of the aequorea Victoria green fluorescent protein. *Science* 1996;**273**:1392–5. https://doi.org/10.1126/science.273.5280.1392.
40. Kinjo AR, Bekker G-J, Suzuki H. *et al.* Protein Data Bank Japan (PDBJ): updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic Acids Res* 2017;**45**:D282–8. https://doi.org/10.1093/nar/gkw962.
41. Singh A, Copeland MM, Kundrotas PJ. *et al.* GRAMM web server for protein docking. *Methods Mol Biol* 2023;**2714**:101–12. https://doi.org/10.1007/978-1-0716-3441-7_5.
42. Krissinel E, Henrick K. Protein interfaces, surfaces and assemblies service Pisa at european bioinformatics institute. *J Mol Biol* 2007;**372**:774–97. https://doi.org/10.1016/j.jmb.2007.05.022.
43. Hu Y, Peng T, Gao L. *et al.* CytoTalk: de novo construction of signal transduction networks using single-cell transcriptomic data. *Sci Adv* 2021;**7**:eabf1356.
44. Jin S, Guerrero-Juarez CF, Zhang L. *et al.* Inference and analysis of cell-cell communication using cellchat. *Nat Commun* 2021;**12**:1088. https://doi.org/10.1038/s41467-021-21246-9.
45. Louault K, Li R-R, DeClerck YA. Cancer-associated fibroblasts: Understanding their heterogeneity. *Cancer* 2020;**12**:3108. https://doi.org/10.3390/cancers12113108.
46. Neefjes J, Jongsma M, Paul P. *et al.* Towards a systems understanding of MHC Class I and MHC Class II antigen presentation. *Nat Rev Immunol* 2011;**11**:823–36. https://doi.org/10.1038/nri3084.
47. Xu Z, Wang W, Yang T. *et al.* STOmicsDB: a comprehensive database for spatial transcriptomics data sharing, analysis and visualization. *Nucleic Acids Res* 2023;**52**:D1053–61. https://doi.org/10.1093/nar/gkad933.

Briefings in Bioinformatics, 2025, **26(1)**, bbae716
https://doi.org/10.1093/bib/bbae716
Problem Solving Protocol