CGCom: a framework for inferring Cell-cell Communication based on Graph Neural Network

# Honglin Wang1, Chenyu Zhang1, Seung-Hyun Hong1, Peter Maye2, David Rowe3, Dong-guk Shin1

1Computer Science and Engineering Department, University of Connecticut, Storrs, CT 06269, USA

2Department of Reconstructive Sciences, Univ. of Connecticut Health Center, Farmington, USA

3Center for Regenerative Medicine and Skeletal Development, Univ. of Connecticut Health Center, Farmington, USA

Corresponding e-mail: honglin.wang@uconn.edu

**Abstract**— Cell-cell communication is crucial in maintaining cellular homeostasis, cell survival and various regulatory relationships among interacting cells. Thanks to recent advances of spatial transcriptomics technologies, we can now explore if and how cells’ proximal information available from spatial transcriptomics datasets can be used to infer cell-cell communication. Here we present a cell-cell communication inference framework, called CGCom, which uses a graph neural network (GNN) to learn communication patterns among interacting cells by combining single-cell spatial transcriptomic datasets with publicly available ligand-receptor information and the molecular regulatory information down-stream of the ligand-receptor signaling. To evaluate the performance of CGCom, we applied it to mouse embryo seqFISH datasets. Our results demonstrate that CGCom can not only accurately infer cell communication between individual cell pairs but also generalize its learning to predict communication between different cell types. We compared the performance of CGCom with two existing methods, CellChat and CellPhoneDB, and our comparative study revealed both common and unique communication patterns from the three approaches.

Commonly found communication patterns include three sets of ligand-receptor communication relationships, one between surface ectoderm cells and spinal cord cells, one between gut tube cells and endothelium, and one between neural crest and endothelium, all of which have already been reported in the literature thus offering credibility of all three methods. However, we hypothesize that CGCom is superior in reducing false positives thanks to its use of cell proximal information and its learning between specific cell pairs rather than between cell types. CGCom is a GNN-based solution that can take advantage of spatially resolved single-cell transcriptomic data in predicting cell-cell communication with a higher accuracy.

Keywords— Graph neural network, Spatial transcriptomics, Cell-cell Communication.

# Introduction

Multicellular organisms rely on the precise coordination of cellular activities, which hinges upon effective intercellular communication among the various cell types and tissues within the organism [1]–[3]. Thus studying cellular functions increasingly requires how cells communicate with each other given a temporal and spatial context. It is known that signaling events are frequently orchestrated through diverse protein interactions such as ligand-receptor, receptor-receptor, and extracellular matrix-receptor interactions. In addition, the recipient cells, in response, initiate downstream signaling via their respective receptors, typically culminating in the modulation of transcription factor activity and subsequent changes in gene expression.

Historically, communication could only be studied via *in vitro* experiments consisting of one or two cell types and a selected number of genes. Thanks to the advances in single-cell technologies and spatial transcriptomics, which measures gene expression at a single cell resolution, scientists can now attempt to identify cell-cell communication as well as cell type communication. Currently there are many computational tools that can infer cell communication from single-cell transcriptomics. CellChat[4] computes cell type communication probabilities using the law of mass action and considering the geometric means of ligand and receptor expressions weighted by their agonists/antagonists. CellPhoneDB[5] generates a list of most statistically significant ligand-receptor interactions by generating null distribution by permuting cell cluster labels. MISTY[6] constructs a gene communication signaling network within and between cell clusters. SpaOTsc[7] and COMMOT [8] generate cell communication matrix for a given signaling pathway by solving an optimal transport problem. Unfortunately, these methods suffer from two main drawbacks: (i) These methods mostly focus on inferring cell communication at the cell group level as opposed to the single cell level. (ii) The existing methods do not adequately exploit the spatial information among cells and prior knowledge about interacting cells, which are crucial for inferring cell communication because cells are unlikely communicate with each other if they are far away from each other and certain cell types of cells tend to have a high or low probability to communicate with one another. This limitation often introduces false-positive communication patterns.

With the continuous advancement of neural networks, the emerging field of graph neural networks (GNNs) has fundamentally transformed our capacity to comprehend relationships among nodes within graphs by harnessing the power of information propagation from neighboring nodes. We see an opportunity to use this advanced form of neural network to infer cellular communication. Specifically, we hypothesize that the physical locations of cells can be modelled into interconnected graph structures with nodes and edges pairing them and the signal transmission between neighboring cells can modelled into strength discerning the degree of cell-cell interaction. Considering that cells of a specific type display distinct gene expression patterns [9], we also postulate that, to a certain extent, leveraging information on ligand and receptor expressions can also serve as a means to identify cell types and cells’ communication likelihood. For example, a cell must exhibit a specific ligand or receptor expression pattern to be classified as a particular cell type. GNNs are capable of updating cell ligand/receptor expression through interactions with nearby cells. By training the model for cell type classification based on ligand/receptor expression values, we can interpret the relationship coefficients between nodes as communication coefficients when deciphering cell-to-cell communication.

Building on this motivation, we present a deep learning framework called CGCom (Cell GNN-based Communication) designed to model cell-to-cell relationships and the intricate communication patterns that underlie them. The framework takes as input a series directed sub-graph generated from cell physical locations, combined with ligand expression values, and utilizes cell type information as the training objective. The paired cell communication coefficient is computed from the attention scores in the well-trained GAT graph classifier. CGCom then introduces a heuristic computational algorithm to quantify communication between neighboring cells through various ligand-receptor pairs. Our framework demonstrates promising performance in cell classification when evaluated on three mouse embryo seqFISH datasets, using “best practice” cell labels from the same tissue, and it outperforms multilayer perceptron (MLP) baseline. To illustrate the power of CGCom in inferring cell communication, we employ the attention scores from GAT classifier to infer cell communication on the same datasets, revealing common communication patterns between the three datasets. We posit that this GNN-based framework is both powerful and flexible, offering significant potential in predicting cell types and inferring cell-to-cell communication.

# Methods

## Overview of CGCom

An overview of the CGCom workflow is given in Fig. 1. This framework consists of three key computational components: data preprocessing modules, the Graph Attention Network (GAT) graph classifier and the cell communication inference module. In the preprocessing module, the CGCom takes the gene expression value, cell label and cell location to build the dataset for the classifier including the training graph and training node feature values. The GAT graph classifier performs deep learning training to acquire cell type information, leveraging a combination of subgraphs generated from cell spatial locations and preprocessed gene expression values for the cells. Subsequently, the cell communication inference tool utilizes the attention scores from the GAT classifier and incorporates gene expression data to infer communication score between pairs of cells.

|  |
| --- |
| **Fig. 1.** *CGCom takes the gene expression matrix, cell type and cell label as the input. The GAT learns the ligand expression patterns of different cell types in a semi-supervised model. The framework then extract the attention score in each graph embedding layer in the GAT from the trained model and infer the communication using a heuristic rule. The communication results finally analyzed at both cell level and cell group level.* |

## Curate ligand-receptor-downstream database

Iintercellular responses are initiated with the binding of a ligand to its corresponding receptor, activating specific cell signaling pathways. The foundation for Cell-Cell Communication (CCC) inference lies in constructing a database of ligand-receptor pairs. CGCom uses this database comprised of two parts: 1) ligand and receptor pairs imported from CellChat and CellPhoneDB via direct downloaded from their websites, and 2) receptor and downstream route pairs extracted from the KEGG pathway. The KEGG pathway driven receptor and downstream route pairs are constructed by utilizing a random walk algorithm that begins at the receptor and ends at any encountered transcription factor of the pathway. This data collection process produces a database containing 5842 ligand-receptor-downstream route triplets, including 774 unique ligands, 708 unique receptors, and 308 unique downstream routes.

## Dataset preprocessing

The dataset preprocessing consists of two primary steps: gene expression preprocessing and cell graph producing. During the gene expression preprocessing phase, the raw scRNA-seq data is filtered and normalized and subject to the quality control process (QC) using Scanpy [10]. The QC includes dealing with gene expression dropouts and mitigating the influence of extreme values. To remedy the high dropout rate inherent in scRNA-seq expression data and to mitigate the influence of extreme values, we use log transformed gene expression of given gene in cell (), after the original expression value () is normalized with the non-zero gene-wise median value () (Eq. 1). Following this step, the downstream route scores for each cell are computed by applying rPAC [11] to the log-transformed expression values. The pathway scoring system rPAC quantifies the downstream routes by evaluating the consistency between the actual gene regulatory pattern and the expected regulatory pattern for all genes within the route. For a cell , the score () of the downstream route () using all transformed gene expression values () is expressed in Eq. 2.

For the cell graph producing, CGCom uses cell locations to construct a directed graph with a hyperparameter *P* designed to limit the maximum communication distance. For the 2-D location of a given cell , the graph builder places a node at the corresponding location within the graph structure (). Once all nodes are incorporated into the graph, the Euclidean distances are calculated for each pair of nodes ( and ) using Eq. 3. We assume that cell communication is restricted to a specific distance range threshold, say *P*. The objective is to introduce directed edges between a pair of nodes in G, if the distance between the nodes is less than the predefined threshold *P*. This cell graph builder uses Algorithm (1) to add directed edges to the graph.

## The architecture of GAT and the training preprocess

The Graph Attention Network (GAT) [12] is a type of neural network designed for tasks involving graph-structured data, such as graph classification, node classification, and link prediction. GAT uses attention mechanisms to weigh the importance of neighboring nodes when aggregating information from the graph. We compute attention scores of each node in the graph for its neighboring nodes, and then combines their features using these scores. The GAT mechanism can be expressed by Eq. 4 where represents the feature vector of node at the th graph embedding layer, represents the neighbors of node , is a weight matrix used for transforming the features of the neighbors, is an activation function (ReLU or Leaky ReLU) applied element-wise to the aggregated information, and represents the attention score for node ’s neighbor . These attention scores are computed based on the features of both nodes and and are normalized across all neighbors using softmax function so that the attention score can be easily comparable as shown in Eq. 5. Here is a learnable parameter associated with the pair of nodes and , and is computed as a function of their features (Eq. 6) where is a learnable weight vector that determines the importance of different aspects of the node features. To simplify the computation step, we use choose ReLU in this equation.

Within this framework, GAT is trained for a graph classification problem. Consequently, the GAT in CGCom is comprised of two components, as illustrated in Fig. 2: the graph embedding part and the graph classification part. The graph embedding part includes three layers, where each layer has the same number of nodes as the input layer. The output of the graph embedding part serves as the input for the graph classification part, which, in turn, includes an output layer with the number of nodes corresponding to the different cell type classes.

In a practical situation, it is hard for humans to identify cell type for all cells given an microscopic image, but the gene expression values can be easily obtained using variant sequencing algorithms. Thus, the classifier is trained in a semi-supervised learning mode, using only a few human-labeled data. Since the GAT is used for the graph classification, its dataset is further generated from the two preprocessed data in the preprocessing step. For each node in graph , a subgraph is built with the nodes that has a link with node () and the edge that connecting them (). These nodes are also known as the neighbor of node . To stimulated how the expression value of a ligand will impact the cell type classification step, each node in will have its feature vector () that only contains all ligands expression values (774 features). The default ratio of training set is 1% and the ratio of validation set is 10%. We use the training set with Adam optimizer and train the model [13] and adjust the learning rate and other hyperparameters with the validation set. To prevent the overfitting, an early stop mechanism is introduced to the training process when the validation loss starts increasing and the validation accuracy starts decreasing.

## The inference of cell communication

After the training iteration stops, we infer the communication with a heuristic rule by intaking the communication coefficient, the ligand and receptor expression value, and corresponding downstream routes in each cell. For each two linked cells and , their communication coefficient () is computed from the average attention score from three graph embedding layer (Eq. 7). Intuitively both expression values of a ligand of the sender cell and a receptor in the receiver cell should be high, and both expression values of a receptor in sender cell and a ligand in the receiver cell should be low if one were to infer paracrine cell communication. We compute the communication score from to through ligand and receptor pair () using Eq. 8 where ligand and receptor pair () expression value in both cells *i* and *j* are denoted by () and (), respectively, and the downstream route () score is denoted by . Below used in Eq. 8 is the function computing the ligand and receptor communication strength as shown in Eq. 9. To scaling down the communication score and prevent all the values from becoming close to zero, we use log1p transformed values instead of using the raw value.

To generalize specific cell-cell communications to cell group-level interactions, we aggregate all the communication scores from group to group by grouping feasible ligand and receptor pairs (). We employ a permutation test to determine the statistical significance of the inferred cell group communication. The permutation test involves random permutations of cell labels followed by the recalculation of the average group communication score between and via the communication ligand and receptor pair (Eq. 10). Here, represents the average communication score for cells in group α communicating with cells in group β through a particular ligand and receptor pair () (Eq. 11). corresponds to the average communication score after the cell labels have been randomly permuted. We repeat this process for a total of random samples, with default set to 100 and the default p-value threshold for the statistically significant communication is set to 0.05.

# Results and discussion

## Experiment materials

Our computational experiments use the datasets published by Lofoff et al. (2022) which were generated by integrating seqFISH datasets from midline sagittal sections on three mouse embryos at the 8-12ss (corresponding to the embryonic day (E)8.5–8.75) with two existing scRNA-seq atlas [14]. The authors’ seqFISH dataset provided spatially resolved expressions of 387 target genes, but they used the comparable scRNA-seq atlas to impute missing genes of the spatially resolved cells and produced an extensive list of mouse embryo cell types. The integrated data was obtained from <https://crukci.shinyapps.io/SpatialMouseAtlas/>. All three integrated embryo datasets, namely, E1, E2 and E3, have 29452 genes and have 17806 cells, 14185 cells and 20577 cells, respectively. All these datasets are published with “best practice” total 23 cell types of cell labels as presented in Fig. 3. The figure also shows the number of cells included in each cell type along with the respective 2D plotting of cells with cell type colored accordingly. Noticeable is the distinctive shape of E2 which could be due to sectioning irregularity performed on their 2nd embryo. Our rendering of this 2D cell type coloring confirms the authors’ suggestion that their 2nd embryo suffers from depletion of mesodermal and tail-specific populations [14].

## CGCom can effectively identify cell types

To assess the performance of the cell GAT classifier in CGCom, we use all three datasets as benchmarks. We explored the cell type prediction performance in two steps. First, we studied if and how the maximum communication ratio (*P*) may affect the classification accuracy. The grid search method is used with *P* ranging from 0.0001 to 0.02. The relationship between accuracy and *P* in all three datasets with 1% human-labeled samples are shown in Fig. 4. As shown, the GAT accuracy increases monotonically when *P* increases in the beginning, but after the knee of the curve (0.005) *P* no longer affects the classification accuracy. In all three datasets, the *P* with the best accuracy lands around 0.005 and we use this value as the best hyperparameter in our subsequent analyses. Noticeable in this performance analysis is that the plot for E2 stands out possibly due to block irregularity mentioned earlier. Nevertheless, the pattern of the knee of the curve does not change providing an added confidence in our way of choosing *P.*

|  |  |
| --- | --- |
| **Fig. 3.** *The details of E1, E2 and E3 datasets. Each datasets have 23 type cells.* | A graph of a gat classifier  Description automatically generated  **Fig.4.** *The relationship of accuracy and the max communication distance with 1%the training sample* |

Second, we conducted a comparative analysis by pitting our cell classifier against an MLP-based classifier using identical learning rates, the same number of hidden layers, and the same optimizer. Four criteria were employed to assess the performance of the classifiers including: Accuracy, Multi-class Precision, Multi-class Recall, and Multi-class F1-score. We varied the ratios of human-labeled samples at 1%, 5%, and 10%. As shown in Table 1, the performance of GAT and MLP increases when the ratio of human-labeled samples increases. Compared to MLP, GAT always outperformed MLP with ~30% improvement in testing Accuracy, and ~0.3 improvement in F1 score, Recall and Precision.

|  |
| --- |
| **Table 1.** *The detailed performance comparing GAT with MLP for all three datasets under different training ration situation when is set to 0.005.* |

Our conclusion is that CGCom's outperformance over MLP is due to the former’s use extra information, that is, the cell type information of neighboring cells which is not used at all in the MLP model. The next question is then what if we instruct the model to pay a deeper “attention” to utilize ligand-receptor relationships embedded in gene expression values on top of the cell graph. Can the model now predict cell-cell communication patterns? This topic is further discussed in the following subsections.Top of Form

## CGCom can identify cell communication

To assess the performance of CGCom in inferring cell communication, we first check communication coefficients between each pair of cell types. i.e., how frequently/infrequently cells of two types communicate with each other. Fig. 5a shows a summary view of such coefficients between a pair of three datasets, E1, E2 and E3, in the form of heatmaps. This heatmap rendering also exhibited the irregularity of E2, as we anticipated. Nevertheless, the analysis was performed on three comparable tissue blocks, their heatmap patterns should be consistent across all three cases and any such case would be considered more likely a true positive. Two cases are notable, one enclosed by small square blue boxes and one enclosed by small square red boxes, in Fig 5A. The first one is for communication coefficient between “Anterior somitic tissues” and “Intermediate mesoderm” which was high in all three cases. The related finding in the literature survey is “Somitic tissues can influence the development of intermediate mesoderm during embryogenesis” by Tani et. al. [15]. How this relationship is revealed in our spatial expression mapped image is shown in Fig. 5B1. The three expanded views enclosed by blue boxes show that somitic tissues cell are neighboring with intermediate mesoderm cells. The second pair for the high communication coefficient is between “Surface ectoderm” and “Anterior somitic tissues”. The related finding in the literature survey is “Surface ectoderm is necessary for the morphogenesis of somites” by Correia et. al. [16]. This relationship is also clearly visible in our spatial expression mapped image shown in Fig. 5B2. In all three expanded views enclosed by red boxes in Fig. 5B2, surface ectoderm cells are neighboring with anterior somitic tissues cells.

|  |
| --- |
| **Fig. 5. *A.*** *The attention score between each cell type pairs in all three datasets.* ***B****. The visualization of cell location in each dataset for the selected cell type pairs. The red box stands for the area that two types of cells are neighboring.* |

We also conducted a comparative analysis with two widely used methods, CellChat and CellPhoneDB. Fig. 6 shows this comparative study outcome in the form of Sankey diagrams as these can succinctly visualize the ligand and receptor communication patterns identified by all three datasets. In Fig. 6A1, each tile shows how ligands listed on the left interact with their corresponding receptors listed on the right. The bandwidth connecting two points in each side of the tile represents the communication score derived using each method. The same comparison analysis was performed on all three datasets, thus generating a total of 9 tiles. An interesting discovery in this comparison is that CGCOM and CellPhoneDB could be more immune to irregular sectioning (i.e., only E2 analysis outcome differs for CellChat). Examining differences in bandwidth also reveals which ligand and receptor pairs are statistically significant and more commonly appearing across different analyses. Common pairs appearing in all three datasets are individually shown in Fig. 6A2, 3 and 4 and they are enclosed by black boxes with genes names stated in them. These three subfigures also include two cases: one that is discovered only by CGCom, enclosed by green boxed (Fig 6A2); and one that is discovered by CellPhoneDB and CellChat, enclosed by red boxes (Fig 6A3 and Fig6A4, respectively). We discuss later that the one discovered by CellPhoneDB and CellChat is more likely a false positive. To further explore the

|  |
| --- |
| **Fig. 6. *A.*** *Sankey diagram for the ligand receptor using all three datasets with three different methods.* ***B.*** *The communication patterns details and visualization including three identified by all the methods.* |

communication patterns, Fig. 6B shows Venn diagrams visualizing how significantly the patterns obtained from the three methods overlap with each other. It shows that a total of 598 ligand-receptor pairs were commonly identified by

all three methods. Regarding pair-wise comparisons, CellPhoneDB vs. Cellchat included a very large proportion patterns (2922) commonly but not so much between CellPhoneDB vs. CGCom and Cellchat vs. CGCom suggesting that CGCom is most discriminating (i.e, least overlapping with outcomes from other two methods).

Fig. 6B also shows five specific ligand-receptor communication pairs (P1 – P5) that we identified. Of these P1-P3 are predicted by all three methods (i.e., one of 598 common communication relationships). P1 involves “surface ectoderm cells” secreting ADCYAP1 (adenylate cyclase activating polypeptide1) ligand which is detected by ADCYP1R1 (ADCYAP receptor type 1) receptor. The work of Xu et al. in 2016 [17] states that ADCYAP1 plays a role in the proliferation of hematopoietic progenitor cells in murine bone marrow. Additionally, Mabuchi et al. in 2004 [18] reports ADCYAP's importance in the development of spinal sensitization and neuropathic pain induction. Our results highlight that surface ectoderm cells may influence spinal cord cells through the ADCYAP1 signal. Fig 6B also includes where in our spatial expression mapped images for E1 – E3, this P1 relationship appears (enclosed by red rectangle boxes). P2 involves “gut tube cells” secreting the apelin receptor early endogenous ligand (APELA), whose signals are received by “endothelial cells” through the apelin receptor (APLNR). Helker et al. in 2020 [19] sates that Apelin signaling propels vascular endothelial cells into a pro-angiogenic state. The cells exhibiting P2 patterns are also captured in its corresponding spatial expression mapped images (enclosed by red rectangle boxes). P3 involves “neural crest cells” communicating with “endothelial cells” through pleiotrophin (PTN) and integrin subunit beta 3 (ITGB3). Heiss et al. [20] report that PTN may mediate the pro-angiogenic effects of endothelial progenitor cells.

|  |
| --- |
| **Fig.7. *A.*** *The visualization of paired neural crest and endothelium cell communication through PTN and ITGB3 in E1 dataset.* ***B.*** *The violin plot for the ligand and receptor expression value in Neural crest cell and endothelium. The orange violin stands for the communicating cell and blue violin stands for the non-communicating cell.* ***C.*** *9 different downstream route scores of ITGB3 in the endothelium.* |

Among the five patterns (P1 – P5), P4 is uniquely reported by CGCom. P4 involves “spinal cord” and “neural crest cells” through Neuregulin family gene (NRG1/NRG3) and erb-b2 receptor tyrosine kinase 2 (ERBB2). Piotrowski et. al. [21] states that spinal cord precursors utilize neural crest cell mechanisms to generate hybrid peripheral myelinating glia and one of the spinal cord cells (motor exit point glia) development depends on atonally derived NRG1. Despite this seemingly supporting relationship reported in the literature, P4 is one of 1464 relationships uniquely identified by CGCom.

Another noticeable pattern is P5 which is suggested between “blood progenitor cells” and brain cells through FGF10 and FGFR1. It belongs to one of 2922 relationships commonly discovered by CellChat and CellPhoneDB but not called by CGCom. The reason behind CGCom not calling this pattern is visible in our spatial expression mapped images for E1 and E3; here blood progenitor cells (enclosed by red boxes) are positioned very far away from Forebrain/Midbrain/ Hindbrain cells (pink colored cells). In both cases, these two cell types are not within the cell proximity used to train the model. In E2, only five blood progenitor cells were present. Is it truly a false positive relationship or not may require further study. Nevertheless, not calling P5 suggests that CGCom's use of cell proximity information can discern proximity-driven communication pattern decision making.

Lastly, we discuss how the discovery of a relationship by CGCom can be explained in terms of gene expression statistics of interacting cell types. Fig. 7A illustrates the discovery of the PTN-ITGB3 communication relationship between “neural crest cells” and “endothelial cells” in E1. Its two large red boxes show expanded views of positional layout of cells (nodes), their types (by color, orange vs. black), and ligand-receptor pairing (edges between nodes). Shown in Fig 7B are the violin plots for gene expression levels of PTN and ITGB3 in both cell types (thus showing four panels of violon plots). Nural crest cells are high in PTN and ITBG3 are high in endothelial cells (gene expression pattern showing (above the red line indicating the mean z-score). The two genes can form a ligand-receptor pair is known along with their corresponding cell types encoded into edges interconnecting nodes. Inferring that the ligand signal from a neural crest cell is passed to the receptor of an endothelial cell is plausible and captured by the red arrow given in Fig 7B. Fig 7C boxplot illustrates another piece of information that is used in the inference. These boxplots show the distribution of pathway scores downstream of the ligand-bound receptor cells (orange colored) and the same for receptor signal weak/absent cells (blue colored), i.e., cells that may not receive ligand signals. It shows that all the ITGB3 downstream scores are higher than 0 (red line) indicating that the downstream pathway routes beyond the receptor could have been all activated. Unfortunately, receptor signal weak/absent cells (blue colored) also show activated downstream signals, suggesting limitations in utilizing the downstream pathway score for the inference because activation signals may originate from different sources—a topic which is notoriously difficult to unravel.

# Conclusion and future work

Inferring cellular communication from spatial transcriptomics data is a challenging problem due to high-volume, high-sparsity, and noisy nature of data. CGCom is a deep learning method aiming to infer cell-cell communication by utilizing the two-dimensional x-y coordinate information available for each cell’s gene expression patterns. Specifically, CGCom exploits each cell’s ligand and receptor signal intensity information that can be combined with cell’s proximity information. By mapping the subcellular data from microscopic images to a graph, CGCom can take advantage of not only the cell’s proximity information but also the relationships between neighboring cells.

We reported a few benefits of CGCom. First, when comparing to the MLP model-based cell communication inference methods, CGCom provides a better performance in cell type classification. Second, CGCom integrates ligand and receptor relationship of neighboring cells into the learned model so that the attention scores at each layer can be translated into communication coefficients providing some degree of interpretability, that is, if nearby cells may or may not communicate with each other even through the ligand and receptor is highly expressed. Third, CGCom can accurately infer cell communication between not only pairs of individual cells but also pairs of cell types.

The current CGCom can be further extended. First, its edge weight scheme between a pair of neighboring cells can be extended beyond the current way of examining the ligand and receptor interaction relationships. For example, cells are known to intimately interact with the environment (i.e., extra cellular matrix) and the environment could fundamentally influence how cells communicate with each other. The current model is amenable to such an extension because incorporating data for the involved molecular elements (e.g., ECM deposited proteins, lipid molecules, etc.) can be easily done. Another direction for extension is enhancing the scoring of each cell’s state upon receiving signals from neighboring cells. Our current route-based pathway scoring to measure the activation/ suppression of receptor downstream activity is still quite limited. There are inherent difficulties in addressing this problem due to the sparse nature of data available from spatial transcriptomic assays. Nevertheless, pathway scoring methods such as rPAC [11] and even crosstalk inference tool such as ctBuilder [22] could offer ways to improve analyzing cellular states at the single cell level and thus enabling one to produce a better cell-cell communication prediction system.

# Code implement and availability

CGCom is implemented using Python 3.10 and Pytorch Geometric [23] and it will made open source and publicly available at GitHub (<https://github.uconn.edu/how17003/CGCom>) with the publication of this manuscript.

##### Acknowledgment

Research reported in this work was supported in part by NIH Grant No. 4U54AR078664-03. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

##### References

[1] C. Bonnans, J. Chou, and Z. Werb, “Remodelling the extracellular matrix in development and disease,” *Nat Rev Mol Cell Biol*, vol. 15, no. 12, pp. 786–801, Dec. 2014, doi: 10.1038/NRM3904.

[2] H. Rouault and V. Hakim, “Different cell fates from cell-cell interactions: core architectures of two-cell bistable networks,” *Biophys J*, vol. 102, no. 3, pp. 417–426, Feb. 2012, doi: 10.1016/J.BPJ.2011.11.4022.

[3] X. Zhou *et al.*, “Circuit Design Features of a Stable Two-Cell System,” *Cell*, vol. 172, no. 4, pp. 744-757.e17, Feb. 2018, doi: 10.1016/J.CELL.2018.01.015.

[4] S. Jin *et al.*, “Inference and analysis of cell-cell communication using CellChat,” *Nature Communications 2021 12:1*, vol. 12, no. 1, pp. 1–20, Feb. 2021, doi: 10.1038/s41467-021-21246-9.

[5] M. Efremova, M. Vento-Tormo, S. A. Teichmann, and R. Vento-Tormo, “CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes,” *Nature Protocols 2020 15:4*, vol. 15, no. 4, pp. 1484–1506, Feb. 2020, doi: 10.1038/s41596-020-0292-x.

[6] J. Tanevski, R. O. R. Flores, A. Gabor, D. Schapiro, and J. Saez-Rodriguez, “Explainable multiview framework for dissecting spatial relationships from highly multiplexed data,” *Genome Biol*, vol. 23, no. 1, pp. 1–31, Dec. 2022, doi: 10.1186/S13059-022-02663-5/FIGURES/7.

[7] Z. Cang and Q. Nie, “Inferring spatial and signaling relationships between cells from single cell transcriptomic data,” *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1–13, Apr. 2020, doi: 10.1038/s41467-020-15968-5.

[8] Z. Cang *et al.*, “nature methods Screening cell-cell communication in spatial transcriptomics via collective optimal transport,” *Nature Methods |*, vol. 20, pp. 218–228, 2023, doi: 10.1038/s41592-022-01728-4.

[9] A. Ianevski, A. K. Giri, and T. Aittokallio, “Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data,” *Nature Communications 2022 13:1*, vol. 13, no. 1, pp. 1–10, Mar. 2022, doi: 10.1038/s41467-022-28803-w.

[10] F. A. Wolf, P. Angerer, and F. J. Theis, “SCANPY: Large-scale single-cell gene expression data analysis,” *Genome Biol*, vol. 19, no. 1, pp. 1–5, Feb. 2018, doi: 10.1186/S13059-017-1382-0/FIGURES/1.

[11] P. Joshi, B. Basso, H. Wang, S. H. Hong, C. Giardina, and D. G. Shin, “rPAC: Route based pathway analysis for cohorts of gene expression data sets,” *Methods*, Oct. 2021, doi: 10.1016/J.YMETH.2021.10.002.

[12] P. Veličkovi´veličkovi´c, G. Cucurull, A. Casanova, A. Romero, P. Lì, and Y. Bengio, “GRAPH ATTENTION NETWORKS”.

[13] D. P. Kingma and J. Lei Ba, “ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION”.

[14] T. Lohoff *et al.*, “Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis,” *Nat Biotechnol*, vol. 40, no. 1, pp. 74–85, 2022, doi: 10.1038/s41587-021-01006-2.

[15] S. Tani, U.-I. Chung, S. Ohba, and H. Hojo, “Understanding paraxial mesoderm development and sclerotome specification for skeletal repair,” *Exp Mol Med*, vol. 52, pp. 1166–1177, 2020, doi: 10.1038/s12276-020-0482-1.

[16] K. M. Correia and R. A. Conlon, “Surface ectoderm is necessary for the morphogenesis of somites”, Accessed: Oct. 29, 2023. [Online]. Available: www.elsevier.com/locate/modo

[17] Z. Xu *et al.*, “Pituitary adenylate cyclase-activating polypeptide (PACAP) contributes to the proliferation of hematopoietic progenitor cells in murine bone marrow via PACAP-specific receptor OPEN,” 2016, doi: 10.1038/srep22373.

[18] T. Mabuchi *et al.*, “Development/Plasticity/Repair Pituitary Adenylate Cyclase-Activating Polypeptide Is Required for the Development of Spinal Sensitization and Induction of Neuropathic Pain,” 2004, doi: 10.1523/JNEUROSCI.0983-04.2004.

[19] C. S. M. Helker *et al.*, “Apelin signaling drives vascular endothelial cells toward a pro-angiogenic state,” *Elife*, vol. 9, p. e55589, Sep. 2020, doi: 10.7554/eLife.55589.

[20] C. Heiss *et al.*, “Pleiotrophin Induces Nitric Oxide Dependent Migration of Endothelial Progenitor Cells”, doi: 10.1002/jcp.21313.

[21] L. Fontenas and S. Kucenas, “Spinal cord precursors utilize neural crest cell mechanisms to generate hybrid peripheral myelinating glia,” *Elife*, vol. 10, p. e64267, 2021, doi: 10.7554/eLife.64267.

[22] H. Wang, P. Joshi, S. H. Hong, D. J. Shin, and D. G. Shin, “CtBuilder: A framework for building pathway crosstalks by combining single cell data with bulk cell data,” *Proceedings - 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021*, pp. 270–273, 2021, doi: 10.1109/BIBM52615.2021.9669811.

[23] M. Fey and J. E. Lenssen, “FAST GRAPH REPRESENTATION LEARNING WITH PYTORCH GEOMETRIC”, Accessed: Nov. 01, 2023. [Online]. Available: https://github.com/rusty1s/pytorch\_geometric