GraphComm: A Graph-based Deep Learning Method to Predict Cell-Cell Communication in single-cell RNAseq data

Emily So[1,2,5], Sikander Hayat[3], Sisira Kadambat Nair[1], Bo Wang[4,5,6,7,$],Benjamin Haibe-Kains[1,2,4,6,8,9,$]

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, Canada
[2]Department of Medical Biophysics, University of Toronto, Toronto, Canada
[3]Institute of Experimental Medicine and Systems Biology, UniKlinik RWTH Aachen, Aachen, Germany
[4]Vector Institute for Artificial Intelligence, Toronto, Canada
[5]Peter Munk Cardiac Centre, Toronto, Canada
[6]Department of Computer Science, University of Toronto, Toronto, Canada
[7]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada
[8]Ontario Institute for Cancer Research, Toronto, Canada
[9]Department of Biostatistics, Dalla Lana School of Public Health, Toronto, Canada

$ Co-corresponding authors

Abstract

Cell-cell interactions coordinate various functions across cell-types in health and disease. Novel single-cell techniques allow us to investigate cellular crosstalk at single-cell resolution. Cell-cell interactions are mediated by underlying gene-gene networks, however most current methods are unable to account for complex inter-connections within the cell as well as incorporate the effect of pathway and protein complexes on interactions. Therefore, to utilise relations of cells to ligands and receptors as well as multiple annotations, we present GraphComm - a new graph-based deep learning method for predicting cell-cell communication in single-cell RNAseq datasets. By learning off of a prior model and fine-tuning a network on single-cell transcriptomic data, GraphComm is able to predict cell-cell communication

(CCC) activity across cells, and its impact on downstream pathways, spatially adjacent cells and changes due to drug perturbations.

Keywords: Cell-Cell Interactions, Single-cell RNA, Deep Learning, Graph Neural Networks, Graph Attention Networks

Introduction

Understanding multicellular organisms and their ability to function requires an in-depth understanding of cellular activities[1,2]. This can be achieved by studying the signalling events that will induce responses and downstream effects. The most commonly studied signalling events typically involve the binding of a secreted ligand to a cognate receptor, either in an intercellular (paracrine)[3] or intracellular (autocrine)[4] fashion. The importance of these ligand-receptor interactions in biology has led to an increasing interest in computationally uncovering patterns of cell-cell communication (CCC)[5] and resulting phenotypic effects in healthy, perturbed, and disease conditions. Subsequently, changes in cell-cell communication will be helpful in improving our understanding of tissue function[5] and disease progression[5,6]. Direct cell-cell communication has been shown to be crucial for tumour progression, which enables the transfer of cellular cargo from non-cancerous to cancerous cells[7,8]. Optimal computational tools to model this type of cell-cell communication in tumours can also aid in identifying key patient-selection biomarkers[9] and improving therapeutic approaches for drug response[10].

One of the major challenges in determining cell-cell communication is the lack of gold-standard ground truth in defining potential ligand-receptor pairs, which limits the subsequent validation[11]. Identifying interactions in their native microenvironments often requires expensive experiments or extensive domain knowledge in detecting active interactions[12]. To allow for a more accessible method for studying CCC, single-cell transcriptomics has been incorporated in utilising evidence of gene expression to support CCC activity. As single-cell RNAseq is increasingly used to study cell types and states, there is an imminent

need for computational methods that can perform prediction of validated ligand-receptor activity with count matrices of gene expression. Methods such as CellPhoneDB[13], Crosstalkr[14], NicheNet[15] and CellChat[16], have paved the way for developing cell-cell communication methods that can generate results at both the bulk and single-cell level. These methods have provided interesting results in novel CCC activity within scRNA transcriptomic data. Based on the results from these methods, there has also been increasing interest in applying CCC methods to different modalities such as spatial transcriptomics data, where cell coordinates could provide a basis for visualising CCC predictions with respect to spatial adjacency.

However, there are still multiple limitations that must be addressed to improve the accuracy and biological relevance of CCC prediction. Currently, there is a dearth of a consistent ground truth for validated ligand-receptor (LR) pairs, with accompanying annotation such as protein complex information and pathway information. Moreover, previous methods in CCC use LR databases only as an input source, but do not explicitly exploit them actively in the prediction process. This might lead to the potential loss of important indications of cell-cell communication activity reflected by a change in transcriptomic values. Therefore, we hypothesise that CCC methods can be improved by incorporating ground truth information into the method prediction process itself. For example, LR information can be used in conjunction with new, context specific transcriptomic data that will influence the probability of interactions based on expression values. To provide the ability of processing multiple annotations for a large input dataset at once and subsequently project to a small test set, deep learning can be introduced as a potential solution for a new CCC method. By introducing this new format of CCC prediction framework, novel ligand-receptor interactions can be identified that are indicative of a new cell role or cell function in different contexts.

To address these issues, we present GraphComm, a new Graph-based deep learning method to predict cell-cell Communication from single-cell RNAseq data. In previous methods, there are two separate

levels of view from which CCC results can be interpreted - i) as a set of pairs between ligands and receptor pairs, or ii) a set of pairs between designated cell-type groups. GraphComm utilises ligand-receptor annotation (such as protein complex and pathway information) as well as expression values to construct cell interaction networks using feed-forward mechanisms to learn an optimal data representation and predict probability of CCC interactions. GraphComm is able to assign numerical values that represent relationships between cell groups, ligands and receptors. Therefore, for any given ligand-receptor link, a communication probability based on these embeddings can be extracted and used to rank CCC activity. This comprehensive deep learning framework allows for rich information regarding both cellular signalling networks and transcriptomic information to be captured in predictions and enables GraphComm to prioritise multiple interactions at once. Using single-cell transcriptomics data from *Drosophila*, human hearts and cancer cell lines, we demonstrate the utility of GraphComm in 1) identifying downstream activated pathways, 2) identifying biologically relevant changes in CCC in cancer cells after drug perturbation, and 3) predicting cell-cell interactions in spatial microenvironments.

Results

GraphComm overview

Using a curated ligand-receptor database and single-cell transcriptomic data, GraphComm utilises directed graph representations of cell social networks to infer meaningful, validated and novel ligand-receptor interactions between two cell types or cell clusters **(Figure 1)**. The prediction pipeline of GraphComm is split into two distinct steps: 1) extracting features from a prior model built on ground truth **(Figure 1A),** and 2) using transcriptomic information to predict cell-cell communication present in a single-cell dataset **(Figure 1B)**.

Feature Representation Learning of the Ground Truth

Utilising features for validated ligand-receptor pairs from ground truth is critical for inferring cell-cell communication in transcriptomic data to prioritise cellular processes such as co-expression. This requires a comprehensive, large curated ligand-receptor database, for which we currently use OmniPath[17]. In addition to providing more than 30,000 validated intracellular interactions and more than 3,000 validated intercellular interactions, OmniPath is accompanied by protein complex and pathway information[17,18,19]. Incorporating this information has proven to be successful previously in methods such as LIANA[11] and CellChat[16], incorporating factors on gene expression from protein sub-unit/pathway and extracting more complex patterns from CCC. First, GraphComm utilises an input single-cell RNAseq expression matrix and identifies all significantly expressed ligands and receptors present in the dataset. A directed graph using these ligands and receptors is constructed, with edges drawn from ligand to receptor only if the link occurs in the OmniPath Database. In addition to the graph structure, GraphComm provides further information on protein complexes and pathways in the form of node embeddings using protein complexes and KEGG pathway annotation. For any given pair of two ligand/receptor nodes, there are corresponding numerical values that represent two given ligand/receptors' correlations to pathway and complex similarity (see "Methods" Section). Using this updated directed graph, Feature Representation Learning is conducted via the Node2Vec[20] framework, which will update the numerical information for each node depending on the validated links that exist. Once Representation Learning is completed, the processed outputs of this task are used as downstream input features for validated ligand-receptor links that can be used to further predict CCC activity **(Figure 1A).**

Calculating Communication Probability with RNAseq

After numerical features have been extracted using context-independent information from the OmniPath Database, communication probability is calculated for all possible ligand-receptor links. This calculation begins with the input of a given scRNAseq dataset. GraphComm begins by identifying defined cell groups within the dataset (such as cell types obtained after Louvain/Leiden[21] clustering) and, for each

group, identifying differentially expressed ligands and receptors. A new directed graph is constructed using nodes of three types - cell groups/clusters, ligands and receptors. Edges are drawn from a given cell group to a ligand/receptor if the expression of the ligand/receptor in the cell group is significantly up-regulated. This newly created graph resembling the potential cellular network is further annotated by a combination of previously obtained positional embeddings and new contextual information (see "Methods" Section). The annotated directed graph is fed as input to a Graph Attention Network (GAT)[22] for a training period. Numerical values for all nodes are continuously updated, skewing towards a binary ground truth favouring interacting ligands and receptors. Finally, a table containing values for all possible ligand-receptor links in the scRNA dataset are obtained via inner product computing of the GAT output, which are used as communication probability **(Figure 1B).**

CCC results and visualisation

To visualise possible ligand-receptor links in a dataset, the new matrix from the last step of the GraphComm framework can be utilised in combination with the input graph **(Figure 1C)**. For each possible ligand/receptor pair in the original input graph, a new link is created with a corresponding weight value from the communication probability matrix. This allows for ranking of ligand-receptor pairs for inference and visualisation, as well as identifying source and destination cell groups.

Predicting pathway activity observed in virtual *in situ* hybridizations

To demonstrate GraphComm's ability to validate spatial communication patterns found in in-situ hybridizations, we studied pathway activity from CCC in a scRNA dataset of a *Drosophila melanogaster* embryo at developmental stage 6[23] **(Figure 2A)**. In the original publication, upon querying expressed ligands and receptors in the spatial reconstruction, the authors identified patterned expression of the Hippo signalling pathway in stage 6 *Drosophila* embryo. There was also further correlation found with Hippo pathway activity[24,25] and mitotic onset/cell cycle processes, providing insight into signalling previously unfound in early *Drosophila* development.

GraphComm analysis on the stage 6 *Drosophila* embryo identified activity involving the Hippo pathway activity in top-ranked interactions. Inference was conducted on a cumulative cell-gene matrix on the six scRNA datasets of stage 6 *Drosophila* embryos **(Figure 2B)**. Among CCC interactions in which a ligand/receptor is a member of the Hippo pathway, there is a strong presence of other pathways including animal mitophagy[26], MAPK[27] signalling, cell cycle processes[28], Phospholipase D signalling[29] and Notch[30] signalling pathways. Currently in Graphcomm's results of cell-cell communication, the downstream effects of the ligand-receptor interaction cannot be identified. The cell cycle - Hippo signalling pathway interaction found by GraphComm could imply the same correlation as found by the original authors. While not identified in the original publication, MAPK[31],Phospholipase D[29] and Notch [30] signalling were also found to be involved in Drosophila embryonic development previously. Within the top ranked interactions of the prediction set, GraphComm achieved significant coverage of interactions involving Hippo pathway genes, achieving 100% coverage within the top 100 interactions. Coverage from the ordered prediction set was significantly better than 100 iterations of randomization, which averaged a coverage of 44.7% **(Figure 2C)**.

Graphcomm's coverage of the Hippo pathway in top-ranked interactions was then compared to other CCC methods **(Figure 2D)**. All 3 tested CCC methods (CellPhone[13], Connectome[32], CellChat[16]) did not capture any presence of Hippo signalling within the top 1000 interactions. A significant factor of this finding is that only intercellular interactions are utilised in other CCC methods, while GraphComm trains on all both intracellular and intercellular interactions. While there were identified interactions related to the ErbB pathway, which has been found to engage in Hippo crosstalk in *Drosophila*[33], the findings imply that the GraphComm framework was uniquely able to capture previously published results in cell-cell communication.

Predicting changes in cell-cell communication due to genetic and chemical perturbations

To understand the potential of GraphComm in finding pathways and cell-cell communication affected by drug treatment, network inference was conducted on scRNA cancer cell lines pre and post drug treatment. The dataset consists of a pre-post treatment set of PC9 lung adenocarcinoma cell lines[34], with cell populations at Day 0, 3, 7 and 14 obtained after being treated with Osimertinib (a tyrosine Kinase inhibitor[35]). GraphComm was applied to this dataset to determine how the framework could - i) identify more unique interactions in two datasets with different conditions than two biological replicates and ii) identify more common interactions in two biological replicates than two datasets of different condition. Results from GraphComm could be then further used to identify how drug targets/mechanisms of action can be captured in cell cell communication[36]. In this task, the ground truth is still limited to the information we had available in our previous benchmarks - validated Ligand-Receptor (LR) interactions in the OmniPath database and known pathway information annotated via KEGG[18]. GraphComm obtained results on 3 separate datasets across 2 conditions - 1 scRNA dataset of day 0, and two biological replicate scRNA datasets at day 7 **(Figure 3A)**.

We first looked at the top 100 validated intercell interactions in all 3 datasets to identify overlap between biological replicates and different condition datasets. A smaller subset was taken in this benchmark, as the number of intercell interactions in the ground truth are quite small (3599 pairs). There was a 23% overlap between the two post-treatment biological replicates Although a small overlap, the number of common interactions between biological replicates was larger than the 9% overlap between pre and post treatment datasets **(Figure 3B, 3C)**. This difference allows for the discovery of new interactions within the respective datasets that could potentially allude to new findings in PC9 sensitivity or chemoresistance[37]. *TNF*,a multifunctional proinflammatory cytokine that belongs to the tumour necrosis factor (TNF) superfamily[38], was found to be a top-ranked (rank 2) ligand in the day 7 dataset. In previous scientific

studies, it has been found that TNF has high expression in lung cancer and potentially plays an oncogenic role[39] . While the overlap between common top-ranked interactions that GraphComm found was quite low, at 23%, it was a better performance than biological replicate overlap achieved within the top 100 interactions of 100 randomised predictions (**Figure 3D)**. In comparison to results from other established CCC methods, GraphComm prevailed as the method more successful at identifying interactions unique to pre and post treatment, achieving a significantly larger fraction of unique interactions within the top 100 validated intercell interactions. (**Figure 3E).** Therefore, the results from GraphComm imply that our framework allows for identifying unique and potentially biologically meaningful specific interactions in pre and post drug treatment cancer datasets.

Predicting cell-cell communications in spatially adjacent cells

To understand how GraphComm can analyse new directions in CCC, its ability to predict interactions in accordance with other modalities can be assessed. For this inference benchmark, GraphComm is applied on spatial transcriptomics data obtained from human hearts post myocardial infarction, where different regions of the impacted hearts (Ischemic, remote, border and fibrotic zones) affected by myocardial infarction and control[40] (unaffected) were sequenced. The overall objective of this benchmark is to demonstrate GraphComm's ability to learn graph structure in agreement with its spatial microenvironment i.e. cell positioning and adjacency. This unlocks a new dimension to compare against ground truth for the CCC predicted. For all regions of the heart sequenced, cell expression is also accompanied by spatial coordinates. In addition to validating the expected results of ligand-receptor pairs, identifying the proximity of potential source and destination cell groups responsible for that interaction can further influence the feasibility of that interaction **(Figure 4A).**

We applied GraphComm on a single spatial transcriptomics dataset sample human heart obtained from the fibrotic zone (FZ) for CCC inference. The dataset was accompanied by spatial coordinates for all cells,

which could be used to construct a map of the tissue organisation by cell type (**Figure 4B).** The constructed spatial map could then be used to subset for two interacting cell types found via GraphComm and visualise their spatial adjacency**.** When identifying validated intercellular interactions within the prediction set, a top-ranked (rank 6) interaction found by GraphComm includes interactions between vascular smooth muscle cells (vSMCs) and myeloid cells, whose crosstalk and relation has been previously studied in regions of the carotid artery[41]**(Figure 4C)**. The interaction taking place is between *CXCL12*[42] , a stromal cell-derived alpha chemokine ligand and *CXCR4*[43], a stromal cell-derived factor-1 receptor. In previous scientific studies, this interaction has been analysed multiple times specifically for understanding its role in myocardial infarction[44–46], demonstrating high biological relevance to the dataset. In comparison, a low ranked interaction found is between pericyte and neuronal cells, whose spatial adjacency is significantly more sparse **(Figure 4D)**. The ligand-receptor interaction between these two cell groups is between *RELN*[47], a secreted extracellular matrix protein and *ITGB1*[48], an integrin beta subunit. This validated interaction has primarily been observed in glioblastoma cells[49] and is not as relevant for this cell type interaction. In comparison to results from 100 randomization iterations and other CCC methods, the original prediction set of GraphComm had a lower median spatial adjacency of its 10 top ranked interacting cell types across all interactions **(Figure 4E).** These results show that with the correct information provided, GraphComm can not only prioritise biologically important and validated interactions but also prioritise interactions in spatially adjacent cell types.

Discussion

The study of cell-cell communication has been found to provide novel insights into areas of biomedicine that were previously unexplored. By studying ligand-receptor interactions in cells of various cell types, development stage, disease condition or perturbation, critical processes involved in cell-state identity and function can be revealed. While many methods can predict possible CCC activity between cells, using a

ground truth ligand-receptor database and transcriptomic data, these early methods have been found to not fully capture cell social networks as well as be limited in the incorporation of other modalities[5,11]. These initial results call for a need for a more comprehensive method for predicting CCC that captures the complex nature of cell-cell interaction networks, and includes prior ligand-receptor information in CCC prediction.

In this study, we introduce a new graph-based deep learning method, GraphComm, for inferring validated and novel cell-cell communication (CCC) ligand-receptor pairs in single-cell transcriptomic data. GraphComm is able to capture directed graphs demonstrating the relation between cell types and identified ligand/receptors as well as information for prioritising cell-cell interactions, and account for cell positioning, pathway annotation and protein complexes. We include both inter-cell and intracellular Protein-Protein Interactions (PPIs) obtained from OmniPath in our training phase. To account for CCC occurring both across and within one cell/cell type, GraphComm is trained on both intracellular and intracellular PPIs obtained from the OmniPath database so that GraphComm can be used for predicting CCC across and within one cell-type. For evaluation in the three use cases presented in this study, we validated using intracellular interactions for the *Drosophila* study so as to identify results that could reflect the original author's outcome. For the other two datasets, we focus on intercellular communications to reflect the definition of cell cell communication used in previous studies[11,16] and focus on more diverse patterns that are occurring between different cell types/groups. We were able to prove that GraphComm can reproduce CCC results seen through spatial reconstruction, identify unique interactions in cell populations after drug treatment and prioritise interactions occurring in spatially adjacent cell types.

While GraphComm is able to produce some exciting results in identifying CCC interactions across multiple different datasets, there are still some unresolved challenges that exist in the field which require further investigation. In any given dataset, GraphComm identified on average 3000 ligand candidates and

3000 receptor candidates for interaction. Of all the possible interactions that could exist, about 0.2% on average have been previously validated. Due to the nature of limited validated CCC activity, GraphComm and other methods in this field must combat a large class imbalance in learning social networks and filtering out the ground truth in a new dataset.

Also, further exacerbated by the lack of ground truth, it has been previously recognized that there is a lack of gold standard for benchmarking in CCC with other methods. As research continues to develop and more validated CCC interactions are found in different contexts, predictions of ligand-receptor interactions and their evaluation will improve as well. In the presence of those events, we anticipate that GraphComm will continue to be a useful graph-based deep learning method that can accurately capture ligand-receptor events in single-cell transcriptomic data.

Methods

Datasets

Single cell RNA datasets

GraphComm was benchmarked on three different datasets covering different biological settings - developmental, perturbation, and fibrosis (**Table 1)**. The Drosophilla developmental dataset consisted of 5835 cells, and cell clustering was executed using the Leiden[21] algorithm (no pre-existing cell types were set). The dataset can be accessed via the GEO accession code GSE95025. PC9 lung cancer cell line perturbation data consists of 6778 cells for Day 0, 4770 cells for Day 7 (replicate 1), and 3515 cells for Day 7 (replicate 2). Cell clustering was executed using the Leiden[21] algorithm (no pre-existing cell types were set). The dataset can be accessed via the GEO accession code GSE150949. The 10X Visium data consists of one Human Myocardial Infarction spatial slide from the left ventricle of the heart. The slide

dataset consists of 3100 cells, from 9 major cell types.The dataset can be accessed via Zenodo (see "Research Reproducibility" Section).

Curated ligand-receptor database

In selecting a database that could be used to establish a ground truth for cell-cell communication, we have decided to source all validated ligand-receptor interactions from the OmniPath database. Specifically, all scientifically validated PPIs (31729 pairs), intercellular interactions (3599 pairs) protein complex grouping (8022 complexes), and KEGG pathway annotation (7534 entries) were extracted from the OmniPath database via the python package *omnipath* 1.0.6[17,50]. Data extracted using this python package was provided in a series of tables, using HGCN symbols for ligands and receptors as identifiers.

Preprocessing of single-cell data and creation of a directed cell to LR graph

To prepare single-cell RNAseq data into a computable input for predicting cell-cell communication, raw counts were converted to a log-transformed normalised count. Significantly expressed genes were identified for each cell group via identifying genes that had an average non-negative expression across all cells of a given cell type. For each cell group, the set of genes was subsetted to strictly include HGNC identifiers that also appear in the OmniPath database as potential ligand/receptor. Then a directed graph was constructed (using Pytorch Geometric[51]) pointing from each cell group to their respective significant ligands and receptors.

Model training and inference

CCC predictions via GraphComm are generated via a two step pipeline: first building a prior model using the OmniPath Database, and then performing inference using transcriptomics data.

Building of prior model using OmniPath database

After identifying ligands and receptors present in the scRNA dataset based on their expression profile, the OmniPath database is cross referenced to identify validated ligand-receptor present within the set. A directed graph is constructed using the Pytorch Geometric[51] library with ligands and receptors as nodes, and validated edges constructed from ligand to receptor. The directed graph is accompanied by a matrix with the dimensions (# of nodes) x (# of nodes). For each cell in the matrix, which points to the intersection of any two nodes, the value is as follows:

$$V(i,j) = c_{i,j} + p_{i,j}$$

Where $V(i,j)$ is the value for that cell in the feature matrix, $c_{i,j}$ is the number of protein complexes that node i and j are both subunits in and $p_{i,j}$ are the number of KEGG pathways that i and j are both members of. This annotated feature matrix is then trained for 100 epochs through Random Walk via a Node2Vec[20] Model, with an output embedding dimension of 2. The output after the training period is two dimensional positional information for each node in the input graph. Inner products of nodes are used to construct a positional matrix with the shape of number of ligands x number of receptors

Inference on cell-cell communication using transcriptomic data and Graph Attention Network

Determining the ranking of ligand-receptor pairs begins with the directed graph pointing cell groups to ligands and receptors, as constructed in the preprocessing step. Similar to the graph used in building a prior model, before inference, a feature matrix of shape *(# of nodes) x (# of nodes)* is constructed to further annotate the directed graph. The value for each cell, which points to the intersection of any two nodes**,** are as follows for node *i* and node *j:*

If node *i* is a ligand/receptor and/or node *j* is a ligand/receptor:

$$F(i,j) = E_i E_j + P_{ij}$$

Where E_i is the average expression of ligand/receptor *i*, E_j is the average expression of ligand/receptor *j* and $P_{i,j}$ is the value corresponding to the matrix generated by the previous prior model step.

If nodes *i* is a cell group and node *j* is a ligand/receptor:

$$F(i,j) = E_{ij}$$

Where E_{ij} is the average expression of ligand/receptor *j* across cells of cell group *i*.

If nodes *i* and *j* are both cell types, embedding is only provided in the event that there is cell-specific information accompanying the dataset that is indicative of positional information. An example of this could be spatial coordinates for cells.In that case, the embedding value is the following:

$$F(i,j) = C(i,j)$$

Where *C(i,j)* is a function of choosing that resembles the relation between cell type *i* and *j* from positional information. For spatial coordinates, for example, *C(i,j)* is computed as the minimum euclidean distance between any cell of type *i* and cell of type *j*.

The directed, annotated graph is then passed through a Graph Attention Network (GAT)[22], trained for 100 epochs against a binary ground truth. Loss function used is negative log likelihood loss[52]. After the training period, a single forward pass is done to receive two-dimensional embeddings for all nodes in the graph. To obtain top ranked validated LR links, an inner product is done of the ligand embeddings and receptor embeddings, returning a matrix of the shape *(# of ligands) x (# of receptors)*. This matrix is then flattened and ranking of LR links is decided by sorting the inner products in descending order.

To obtain source cell and destination cell types, an inner product is computed of the ligand and receptor embeddings by the cell type embeddings. This returns a matrix of the shape *(# of ligands and receptors*,

assigning communication probability for every possible edge. Interactions are ranked by inner product value.

Randomisation Experiments

The randomization experiments on a given scRNA dataset are done via performing the same method of communication modelling, with the same values for annotations for graphs, but by randomly shuffling the edges of both the OmniPath and transcriptomic graph**.** This will cause the model to perform inference using a randomised ground truth and with different assignments of cell groups to ligands and receptors. For each dataset, randomization was done for 100 iterations.

Research Reproducibility

The sequenced Stage 6 *Drosophila* embryos can be found via the GEO accession code GSE95025[23]. Pre and Post treatment data of PC9 cell lines treated with osimertinib can be found via the GEO accession code GSE150949[34]. Spatial transcriptomics data[40] was obtained from 10.5281/zenodo.6578047.

Code availability

The code for GraphComm and notebooks used to analyse data presented in this study are provided in the GitHub repository https://github.com/bhklab/GraphComm and in the Code Ocean capsule https://codeocean.com/capsule/8269062/tree/v2

Acknowledgements

We thank D. Forster, H. Cui, H. Maan and I. Smith for their insightful comments.

Disclosures

SH reports funding from Novo Nordisk and Askbio GmbH, and is a co-founder of Sequantrix GmbH.

References

1. Rouault, H. & Hakim, V. Different cell fates from cell-cell interactions: core architectures of two-cell bistable networks. *Biophys. J.* **102**, 417–426 (2012).
2. Zhou, X. *et al.* Circuit Design Features of a Stable Two-Cell System. *Cell* **172**, 744–757.e17 (2018).
3. Handly, L. N., Pilko, A. & Wollman, R. Paracrine communication maximizes cellular response fidelity in wound signaling. *Elife* **4**, e09652 (2015).
4. Graeber, T. G. & Eisenberg, D. Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nat. Genet.* **29**, 295–300 (2001).
5. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88 (2020).
6. Leimkühler, N. B. *et al.* Heterogeneous bone-marrow stromal progenitors drive myelofibrosis via a druggable alarmin axis. *Cell Stem Cell* **28**, 637–652.e8 (2021).
7. Pasquier, J. *et al.* Preferential transfer of mitochondria from endothelial to cancer cells through tunneling nanotubes modulates chemoresistance. *J. Transl. Med.* **11**, 94 (2013).
8. Dominiak, A., Chełstowska, B., Olejarz, W. & Nowicka, G. Communication in the Cancer Microenvironment as a Target for Therapeutic Interventions. *Cancers* **12**, (2020).
9. Chiodoni, C. *et al.* Cell communication and signaling: how to turn bad language into positive one. *J. Exp. Clin. Cancer Res.* **38**, 128 (2019).
10. Liu, Z. *et al.* Using dynamic cell communication improves treatment strategies of breast cancer. *Cancer Cell Int.* **21**, 275 (2021).
11. Dimitrov, D. *et al.* Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat. Commun.* **13**, 1–13 (2022).
12. Rao, V. S., Srinivas, K., Sujini, G. N. & Kumar, G. N. S. Protein-protein interaction detection: methods and analysis. *Int. J. Proteomics* **2014**, 147648 (2014).

13. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB v2.0: Inferring cell-cell communication from combined expression of multi-subunit receptor-ligand complexes. Preprint at https://doi.org/10.1101/680926.
14. Weaver, D. T. & Scott, J. G. Crosstalkr: An open-source R package to facilitate drug target identification. *bioRxiv* (2023) doi:10.1101/2023.03.07.531526.
15. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2019).
16. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. Preprint at https://doi.org/10.1101/2020.07.21.214387.
17. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
18. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2015).
19. Tsitsiridis, G. *et al.* CORUM: the comprehensive resource of mammalian protein complexes–2022. *Nucleic Acids Res.* **51**, D539–D545 (2022).
20. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. *arXiv [cs.SI]* Preprint at http://arxiv.org/abs/1607.00653 (2016).
21. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
22. Veličković, P. *et al.* Graph Attention Networks. *arXiv [stat.ML]* Preprint at http://arxiv.org/abs/1710.10903 (2018).
23. Karaiskos, N. *et al.* The Drosophila embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
24. Oh, H. & Irvine, K. D. In vivo regulation of Yorkie phosphorylation and localization. *Development* **135**, 1081–1088 (2008).
25. Huang, J., Wu, S., Barrera, J., Matthews, K. & Pan, D. The Hippo signaling pathway coordinately

regulates cell proliferation and apoptosis by inactivating Yorkie, the Drosophila Homolog of YAP. *Cell* **122**, 421–434 (2005).

26. Youle, R. J. & Narendra, D. P. Mechanisms of mitophagy. *Nat. Rev. Mol. Cell Biol.* **12**, 9–14 (2011).
27. Shilo, B.-Z. The regulation and functions of MAPK pathways in Drosophila. *Methods* **68**, 151–159 (2014).
28. Vergassola, M., Deneke, V. E. & Di Talia, S. Mitotic waves in the early embryogenesis of Drosophila: Bistability traded for speed. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2165–E2174 (2018).
29. LaLonde, M. *et al.* A role for Phospholipase D in Drosophila embryonic cellularization. *BMC Dev. Biol.* **6**, 60 (2006).
30. Baonza, A. & Garcia-Bellido, A. *Notch* signaling directly controls cell proliferation in the *Drosophila wing* disc. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 2609–2614 (2000).
31. Kim, Y. *et al.* MAPK substrate competition integrates patterning signals in the Drosophila embryo. *Curr. Biol.* **20**, 446–451 (2010).
32. Raredon, M. S. B. *et al.* Computation and visualization of cell–cell signaling topologies in single-cell systems data using Connectome. *Sci. Rep.* **12**, 1–12 (2022).
33. Snigdha, K., Gangwani, K. S., Lapalikar, G. V., Singh, A. & Kango-Singh, M. Hippo Signaling in Cancer: Lessons From Models. *Front Cell Dev Biol* **7**, 85 (2019).
34. Oren, Y. *et al.* Cycling cancer persister cells arise from lineages with distinct programs. *Nature* **596**, 576–582 (2021).
35. Bollinger, M. K., Agnew, A. S. & Mascara, G. P. Osimertinib: A third-generation tyrosine kinase inhibitor for treatment of epidermal growth factor receptor-mutated non-small cell lung cancer with the acquired Thr790Met mutation. *J. Oncol. Pharm. Pract.* **24**, 379–388 (2018).
36. AlMusawi, S., Ahmed, M. & Nateri, A. S. Understanding cell-cell communication and signaling in the colorectal cancer microenvironment. *Clin. Transl. Med.* **11**, e308 (2021).
37. Ramirez, M. *et al.* Diverse drug-resistance mechanisms can emerge from drug-tolerant cancer persister cells. *Nat. Commun.* **7**, 10690 (2016).

38. GeneCards Human Gene Database. TNF Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=TNF.
39. Gong, K. *et al.* Tumor necrosis factor in lung cancer: Complex roles in biology and resistance to treatment. *Neoplasia* **23**, 189–196 (2021).
40. Kuppe, C. *et al.* Spatial multi-omic map of human myocardial infarction. *Nature* **608**, 766–777 (2022).
41. Xue, Y. *et al.* Macrophages regulate vascular smooth muscle cell function during atherosclerosis progression through IL-1β/STAT3 signaling. *Commun Biol* **5**, 1316 (2022).
42. GeneCards Human Gene Database. CXCL12 gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CXCL12.
43. GeneCards Human Gene Database. CXCR4 gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CXCR4&keywords=CXCR4.
44. Liehn, E. A. *et al.* Double-edged role of the CXCL12/CXCR4 axis in experimental myocardial infarction. *J. Am. Coll. Cardiol.* **58**, 2415–2423 (2011).
45. Wang, Y. *et al.* C-X-C Motif Chemokine Receptor 4 Blockade Promotes Tissue Repair After Myocardial Infarction by Enhancing Regulatory T Cell Mobilization and Immune-Regulatory Function. *Circulation* **139**, 1798–1812 (2019).
46. Döring, Y., Pawig, L., Weber, C. & Noels, H. The CXCL12/CXCR4 chemokine ligand/receptor axis in cardiovascular disease. *Front. Physiol.* **5**, 212 (2014).
47. GeneCards Human Gene Database. RELN Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=RELN.
48. GeneCards Human Gene Database. ITGB1 gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=ITGB1&keywords=ITGB1.
49. Schulze, M. *et al.* RELN signaling modulates glioblastoma growth and substrate-dependent migration. *Brain Pathol.* **28**, 695–709 (2018).
50. Omnipath. *PyPI* https://pypi.org/project/omnipath/.

51. Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. in *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019).

52. Yao, H., Zhu, D.-L., Jiang, B. & Yu, P. Negative Log Likelihood Ratio Loss for Deep Neural Network Classification. in *Proceedings of the Future Technologies Conference (FTC) 2019* 276–282 (Springer International Publishing, 2020).

Figures

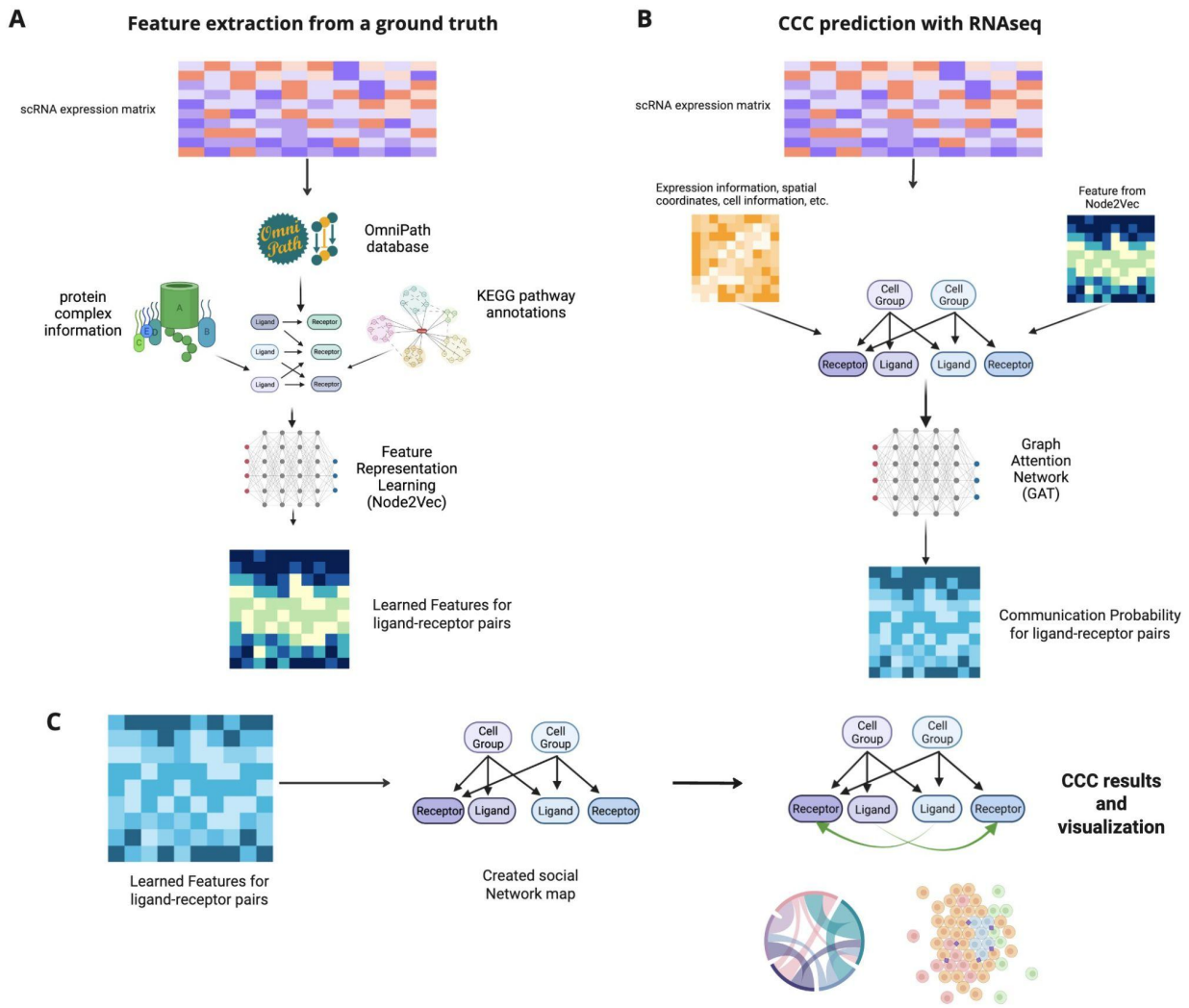

Figure 1: Schematic outlining the architecture of GraphComm to make CCC predictions from scRNAseq data **(A)** GraphComm utilises a scRNAseq dataset with a curated ligand-receptor database to construct a directed graph reflective of the CCC ground truth. Annotating additionally with protein complex information and pathway annotation, Feature Representation is implemented to extract positional information for ligands and receptors within the directed graph **(B)** GraphComm constructs a second directed graph representing the relationship between cell groups and ligands/receptors. Annotating this second directed graph with transcriptomic information, cell group information and positional features from the Feature Representation learning step to receive updated numerical node features via a Graph Attention Network. Node Features can then be used via inner product to compute communication probability for all possible ligand receptor pairs. **(C)** Computed communication probabilities via the Graph Attention Network can be combined with the second directed graph to construct ligand-receptor links with top-ranked CCC activity, which can be used for visualisation of activity at the ligand-receptor and cell group level.

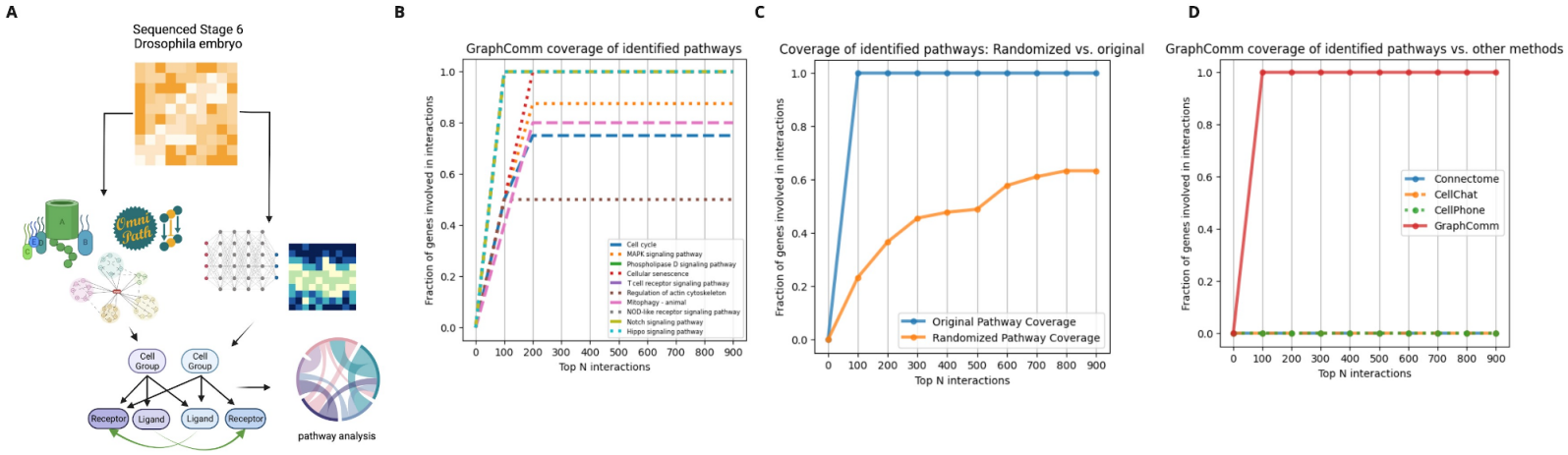

Figure 2: (A) GraphComm architecture can identify previously validated pathway activity from in-situ hybridization experiments. With a scRNA dataset of stage 6 *Drosophila* embryo as input, communication probabilities can be used to identify top-ranked interactions of ligands in the dataset. Corresponding KEGG pathways for ligands and receptors can then be used to visualise pathway activity **(B)** Within the top 1000 interactions, GraphComm was able to achieve 100% coverage of available ligands and receptors participating in the Hippo pathway, which had been found to have patterned expression in *Drosophila embryo*. Hippo ligands/receptors were found to be interacting with pathways that had also been previously studied in *Drosophila* models. **(C)** In comparison to the top 1000 interactions of 100 randomised iterations, GraphComm achieves significantly better coverage of the Hippo pathway using the original directed graph and ground truth **(D)** In comparison to other CCC methods, GraphComm achieves significantly better coverage of the Hippo pathway.

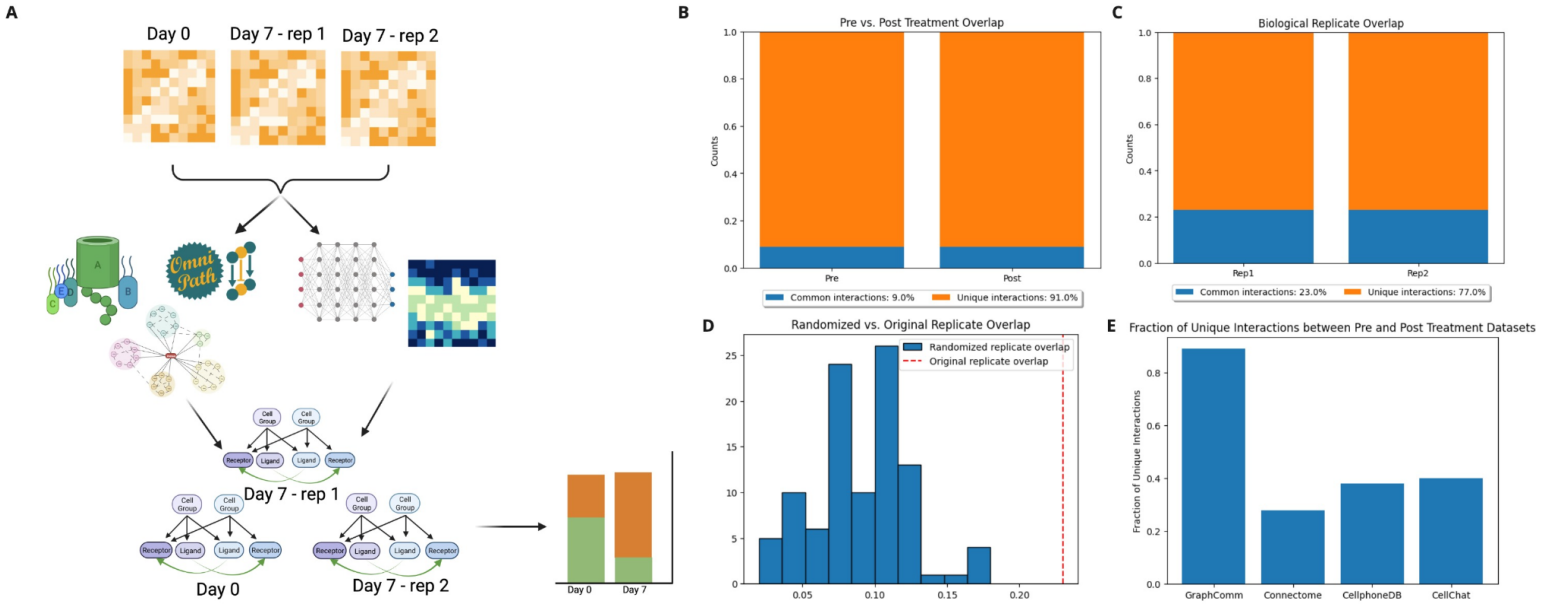

Figure 3: **(A)** GraphComm can identify differences in interactions among datasets that have the potential to be indicative of change in condition. Using multiple datasets of PC9 lung adenocarcinoma cell lines sequenced at different timepoints before and after treatment with osimertinib including two biological replicates, the fraction of overlapping interactions can be compared. This allows for the interpretation of how GraphComm can i) identify more unique interactions in two datasets of different condition than two biological replicates and ii) identify more common interactions in two biological replicates than two datasets of different condition **(B)** Between one PC9 dataset at day 0 (pre-treatment) and one PC9 dataset at day 7 (post-treatment), GraphComm is able to prioritise unique top-ranked interactions, with an overlap of 9% **(C)** In comparison to overlap between pre and post treatment datasets, GraphComm achieves a greater overlap between two day 7 (post treatment) biological replicates, with an overlap of 23% **(D)** In comparison to top-ranked interactions of 100 randomised iterations, GraphComm achieves significantly better overlap between two biological replicated datasets using the original directed graph and ground truth **(E)** In comparison to other CCC methods, GraphComm achieves a higher number of unique interactions between pre and post treatment datasets, potentially indicating GraphComm performs better at identifying interactions that can be indicative of a change in condition.

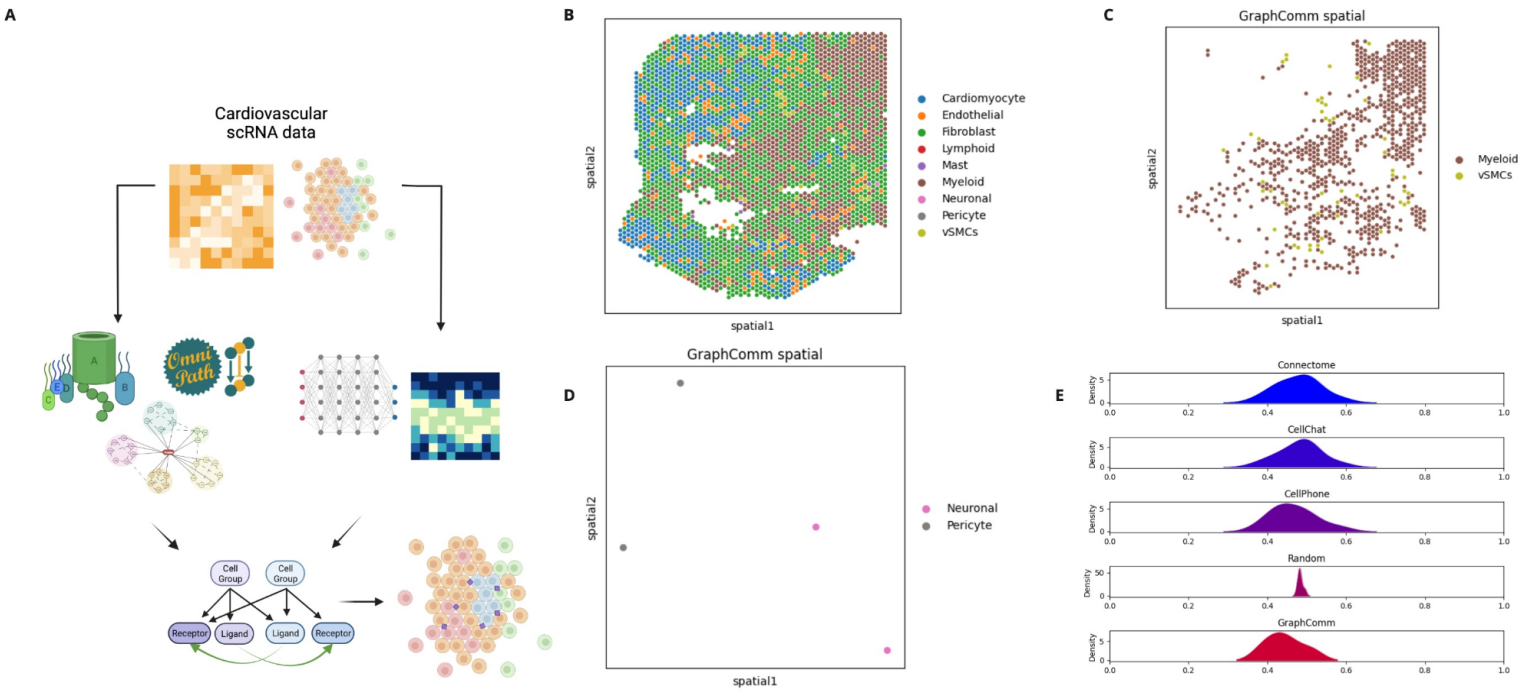

Figure 4: **(A)** GraphComm architecture can predict CCC occurring in spatially adjacent cell groups. Utilising spatial transcriptomics data from the fibrotic region of a patient sample, GraphComm can identify ligand-receptor interactions as well as identify potential source and destination cell groups. Then, the nodes of interest can be used to visualise two interaction cell groups and their respective proximity. **(B)** spatial coordinates of all cells in the dataset and colour coded by the cell-type. **(C)** Spatial map only displaying cells in top interacting cell groups found by GraphComm, Fibroblast and Endothelial cells. From this spatial plot alone, the adjacency of these two cell groups and the potential for interactions can already be observed **(D)** Spatial map only displaying cells in low-ranking interacting cell groups found by GraphComm, Mast and Neuronal cells. From this spatial plot alone, the adjacency of these two cell groups and the potential for interactions can already be observed **(E)** GraphComm had a lower median spatial adjacency of its 10 top ranked interacting cell types in comparison to results from 100 randomization iterations and other CCC methods.

Organism	Tissue type	Modality	Number of cells/spots	Number of Major cell-types	Reference
Stage 6 *Drosophila melanogaster* Embryo	Stage 6 embryo	Single-cell transcriptomics	5835	N/A	Karaiskos et al. 2017[23]
Homo sapiens	Heart Left Ventricle	Spatial transcriptomics	3100	9 cell types	Kuppe et al. 2022[40]
Homo sapiens	PC9 lung cell lines	Single-cell transcriptomics	Day 0: 6778 Day 7 rep1:4770 Day 7 rep2:3515	N/A	Oren et al. 2021[34]

Table 1: Single cell and spatial transcriptomics data used in benchmarking for GraphComm.