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 communication (CCC) is 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. This results in the inability to infer overarching signalling patterns within a dataset as well as limit the ability to successfully explore other data types such as spatial cell dimension. Therefore, to represent transcriptomic data as intricate networks connecting cells to ligands and receptors for relevant cell-cell communication inference as well as incorporating descriptive information independent of gene expression, we present GraphComm - a new graph-based deep learning method for predicting cell-cell communication in single-cell RNAseq datasets. GraphComm

improves CCC inference by capturing detailed information such as cell location and intracellular signalling patterns from a database of more than 30,000 protein interaction pairs. With this framework, GraphComm is able to predict biologically relevant results in datasets previously validated for CCC,datasets that have undergone chemical or genetic perturbations and datasets with spatial cell information.

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], Connectome[15], NicheNet[16] and CellChat[17], 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 and intracellular signalling 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 communication probabilities 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) feature representation learning using a prior model **(Figure 1A),** and 2) using transcriptomic information to predict cell-cell communication present in a single-cell dataset **(Figure 1B)**.

Feature Representation Learning using a Prior Model

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[18]. 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[18,19,20]. Incorporating this information has proven to be successful previously in methods such as LIANA[11] and CellChat[17], 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, receptors and intracellular proteins present in the dataset. A directed graph using these proteins is constructed, with edges drawn from source to target only if the link occurs with validation in the OmniPath Database. Using this updated directed graph, Feature Representation Learning is conducted via the Node2Vec[21] 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 then scaled with a constructed matrix, containing numerical values detailing ligand and receptors' correlation in protein complexes and KEGG pathways (see Methods). These scaled constructed features are used as downstream input features for validated ligand-receptor links that can be used to further predict CCC activity **(Figure 1A).**

scRNAseq infers cell-to-cell communication probability

After numerical features have been extracted using context-independent information from the OmniPath Database, communication probability is calculated for all possible source/target protein links, both intracellular and intercellular. 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[22] clustering) and, for each group, identifying differentially expressed proteins. A new directed graph is constructed using nodes of three types - cell groups/clusters, source proteins and target proteins. Edges are drawn from a given cell group to a source/target protein if the expression of the source/target protein 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)[23] for a training period. Numerical values for all nodes are continuously updated, skewing towards a binary ground truth favouring interacting intercellular ligands and receptors. Finally, a table containing values for all possible source/target protein 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. Additionally, using the chosen datasets GraphComm can perform analyses to confirm previously validated activity and report the potential of new communication (**Figure 2).**

GraphComm predicts Cell-Cell Communication in retrospective datasets

To demonstrate GraphComm's ability to validate spatial communication patterns found in in-situ hybridizations, we studied patterned expression of validated ligands and receptors from CCC in a scRNA dataset of a Drosophila melanogaster embryo at developmental stage 6 **(Figure 3A)**. 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 with activity primarily occurring in the anterior domain. There was also further correlation found with Hippo pathway activity 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 intercell interactions within top-ranked interactions whose patterned expression can be visualised using the package

novoSparc. Utilising this information, the coincidence/possibility of an interaction can be identified. Within the Drosophila embryo dataset, a top ranked intercell interaction was between the secreted ligand *PCNA*[24] and the transmembrane receptor *LAMP1*[25]. *PCNA* has been found to participate in cell cycle interactions, which was found in the original publication to be associated with the patterned expression of Hippo signalling pathway components. *LAMP1*'s role in autophagy can also be connected to Hippo signalling via crosstalk[26], bringing further relevance to the publication. As visualised via the reconstructed embryo, both ligand and receptor are expressed in the same axis (along the posterior region) of the Drosophila embryo, making their interaction feasible and less coincidental given the previous distribution **(Figure 3B,C)**. Alternatively, a lower ranked interaction of the ligand *WNT4*[27] and the receptor *PDE1C*[28] had less overlap, with *PDE1C* only having elevated expression in the posterior region and *WNT4* largely in the anterior **(Figure 3D,E)**. These findings were able to demonstrate that GraphComm can reflect spatial patterns observed in in-situ hybridizations observed in an original publication.

To further assess the ability of GraphComm to identify previously validated inter-cell group communications, we applied our method on an embryonic mouse brain published by Sheikh et. al[29]. In the original publication, the authors were able to identify patterns of communication between identified cell groups including 1,710 unique ligand-receptor interactions that were all validated. The results found in this publication were able to reveal to the author underlying mechanisms and patterns that are critical in mouse brain development and embryogenesis.

The goal of utilising this dataset for CCC inference on GraphComm was to determine, among the top-ranked interactions of the prediction set, how many were also present as an important interactions in the original publication **(Figure 4A)**. In order to determine the overall

relevance and prediction of correctly prioritised interaction members of GraphComm, presence of individual ligands and receptors from validated interactions were compared in the top 100 ranked interactions containing ligand and receptor candidates, in the original test set and the top 100 random iterations. It was found that with the original reference, GraphComm was able to prioritise important ligands and receptors further in the original test set. **(Figure 4B)** While coverage is comparable in benchmarking, the overlap with the publication of GraphComm's top 100 interactions is slightly lower than other methods **(Supplementary Figure 1).** Another version of GraphComm was used for inference comparison as well, trained only on ligands and receptors that were present in the consensus dataset. This training set seemed to reflect the lack of influence of intracellular signalling from the original GraphComm, performing worse than all other methods. The difference in performance can be connected to the difference in prioritisation of certain interactions, with GraphComm prioritising other validated interactions over those explicitly mentioned in the publication. When analysing the full validated interactions that GraphComm covers, it appears that overlap with the original publication converges and performance becomes comparable with other methods, best performing training on the full set of interactions from OmniPath rather than just the intercellular interactions **(Figure 4C).** Therefore, GraphComm was able to identify previously validated ligand and receptor comparably with other methods and although identifying less validated interactions in a top ranked set, could possibly lead to the highlighting of novel activity within the embryonic mouse brain.

GraphComm predicts 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[30], with cell populations at Day 0, 3, 7 and 14 obtained after being treated with Osimertinib (a tyrosine Kinase inhibitor[31]). We used GraphComm to (*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 conditions. Results from GraphComm could be then further used to identify how drug targets/mechanisms of action can be captured in cell-cell communication[32]. In 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[19]. To identify how GraphComm provides inference on single-cell datasets of different and the same conditions, two sequencing timepoints from a cancer cell line dataset were used. One dataset was sequenced before treatment (day 0) and two biological replicate datasets were sequenced 7 days after treatment **(Figure 5A)**.

We first looked at the top 50 validated intercell interactions in all 3 datasets to identify overlap between biological replicates and different conditions. There was a 56% overlap between the two post-treatment biological replicates. The number of common interactions between biological replicates was larger than the 16% overlap between pre and post treatment datasets **(Figure 5B,C)**. To test the significance of the overlap between biological replicates versus pre- and post-

treatment datasets, we conducted 100 randomised iterations of the analysed datasets and were unable to find the same difference in overlap as the original test set ($p < 0.01$) (**Figure 5D)**. 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[33]. The secreted ligand *TGFB1*[34], was found to be a top-ranked (rank 0) ligand in a replicate from the Day-7 dataset. In previous studies, the *TGFB1* protein has been linked to NSCLC prognosis[35] and acquired osimertinib resistance[36] . While the overlap between common top-ranked interactions was not complete at 100%, such level of overlap was greater than overlap observed after conducting inference on 100 randomised iterations ($p < 0.01$) (**Figure 5E)**. In comparison to results from other established CCC methods, GraphComm prevailed as the method most 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 5F).** 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.

To evaluate whether GraphComm can identify CCC leading to drug resistance, we ran our method on a pre- and post-treatment scRNA dataset from small cell lung cancer (SCLC) circulating tumour cell (CTC) derived xenografts[37]. Here, platinum sensitive models were treatment naive, and platinum resistant models were models with relapse following platinum chemotherapy. To investigate the contribution of heterogeneity to therapeutic resistance, the original authors generated xenograft models that were classed as platinum sensitive (treatment-naive) and platinum resistant (already undergone platinum chemotherapy). Both model types were then subjected to treatment of cisplatin and then sequenced, with platinum-sensitive models being subjected to extended treatment until relapse occurred. We subjected one pair of pre-treatment platinum sensitive/platinum resistant models, and one post-treatment pair of

platinum-sensitive/platinum-resistant models to GraphComm for CCC inference. Between the pre-treatment datasets, differences in prioritised ligands and receptors could already be detected. Within the pre-treatment naive model, GraphComm prioritises interactions involving the ligand *ANXA2*[38]. While not specifically identified previously to be contributing to SCLC prognosis and tumour progression, *ANXA2* has been found to have a prognostic role in other types of lung cancer[39] making it a potential ligand of interest for future SCLC therapeutic research. **(Supplementary Table 1)**. Conversely, top-ranked interactions in the pre-treatment resistant model involve the ligand *YBX1*[40], a cold shock domain protein that has shown to be involved in resistance to some platinum based chemotherapies in lung cancers[41]**(Supplementary Table 2)**. Between the post-treatment datasets, GraphComm was able to identify different Ligand-Receptor pairs relevant for treatment resistance. In the post-treatment naive (sequenced upon the event of relapse) dataset, the top ranked interaction involves the ligand *NRP1*[42]*,* an X-linked actin sequester protein. This protein has previously been found to be linked to lung cancer resistance[43] and linked to epithelial-mesenchymal transition[43] in lung cancer cells, a pathway that was highlighted in the original platinum-naive relapsed model **(Supplementary Table 3)**. In the treatment-resistant dataset, within the top 50 validated interactions there was presence of the transmembrane receptor *MCL1*[44]*,* which has been linked many times to the chemoresistance to cisplatin in lung cancer[45,46] **(Supplementary Table 4)**. These results were able to prove that GraphComm can identify different levels of drug response within xenograft models of various conditions.

GraphComm predicts 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 (MI), where different regions of the impacted hearts (Ischemic, remote, border and fibrotic zones) affected by myocardial infarction and control[47] (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 6A)**.

To assess how GraphComm can capture short- and long-range interactions in different histomorphological regions of the heart, we performed CCC inference on 6 slides of the fibrotic zone and 8 slides of the ischemic zone from patients with myocardial infarction. We found that in all fibrotic zone slides, GraphComm detected a large number of interactions between Myeloid cells and vSMCs cells **(Figure 6B)**, whose spatial adjacency makes CCC probable **(Figure 6C)**. A top ranked valid interaction between these two cell groups was between the ligand *ANXA1*[48], a phospholipid-binding membrane protein and the cell surface receptor *FPR3*[49]. While the specific biological role of *FPR3* is not well studied, *ANXA1* has been identified to have an anti-inflammatory, pro-resolving role post-MI[50]. A lower ranking interaction, between Cycling cells and Mast cells (**Figure 6D)** can be visualised as less spatially adjacent making the interaction less probable. Conversely, overall ischemic slides had a large number of interactions between vSMCs and Endothelial cells **(Figure 6E)**, which is also probable as shown by spatial adjacency **(Figure 6F)**. A corresponding ligand-receptor pair is between the ligand *SLIT3*[51] and the receptor *ROBO4*[52]. This specific interaction has been previously found to be closely involved in angiogenesis, a critical biological process involved in healing post-MI[53–55].Similarly to the fibrotic

slides, less frequent interactions between Adipocyte and Mast cells **(Figure 6G)** is also less spatially adjacent.

We then applied GraphComm on a single spatial transcriptomics dataset sample human heart obtained from the fibrotic zone (FZ) for CCC inference for a closer look at benchmarking GraphComm with spatial transcriptomics data, choosing the slide *FZ-GT-P19*. 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 7A)**. Based on GraphComm's results,when analyzing only validated intercellular interactions, patterns can be observed in dominant cell group interactions **(Figure 7B)**. To first identify what regions of the *FZ_GT_P19* slide could be indicative of potential intercellular activity, non-negative matrix factorization (NMF) was conducted on the slide using the framework LIANA+[56] and visualised to determine the important cell groups for observing CCC **(Figure 7C)**. A top-ranked (rank 3) interaction found by GraphComm includes interactions between Myeloid cells and vSMC cells, two cell groups with a mean spatial adjacency of 0.413 **(Figure 7D)**. This cell group interaction is connected to a validated interaction between ligand *HP*[57] and receptor *CD163*[58]*,* whose crosstalk has been previously linked to prognosis of myocardial infarction and heart disease[59] .This finding is also in accordance with important contributors to intercellular activity found with LIANA+, as the most highlighted slide region from the NMF analysis is a region containing Myeloid cells. In comparison, a low ranked cell group interaction is between Cardiomyocyte and Endothelial cells, whose interaction is less probable with a mean spatial adjacency of 0.48 (**Figure 7E).** In comparison to the NMF results as well, these cell groups are present in regions with lower value from the NMF. Against 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 than all other methods **(Figure 7F).** 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[11,17] 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 1000 source protein candidates and 1000 target protein candidates for interaction. Of all the possible interactions that could exist, a very small fraction of them were 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/ischemia (**Table 1)**. All data used is available through public repositories (see Data Availability 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 (30053 pairs), intercellular interactions (3731 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[18,60]. Data extracted using this python package was provided in a series of tables, using HGNC symbols for intracellular proteins/ligands/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 a potential ligand/receptor or intracellular protein. Then a directed graph was constructed (using Pytorch Geometric[61]) pointing from each cell group to their respective significant intracellular protein/ligand/receptor.

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[61] library with each protein as a node, and validated edges are drawn if a pair of proteins participate in a validated intracellular or intercellular interaction. Alongside the directed graph, a matrix is constructed 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. The constructed directed graph is then trained for 100 epochs through Random Walk via a Node2Vec[21] 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 output node embeddings are used to construct a positional matrix with the shape of number of source proteins x number of target proteins and then multiplied with the

constructed annotation matrix to further scale the output in accordance with protein complex and pathway information.

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

Construction of Cell and Ligand/Receptor embedding matrix

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 source/target protein and/or node *j* is a source/target protein:

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

Where E_i is the average expression of source/target protein *i,* E_j is the average expression of source/target protein *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 source/target protein:

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

Where E_{ij} is the average expression of source/target protein *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.*

Model Architecture

We then passed the directed graph from the preprocessing step, with the corresponding embedding matrix, through a Graph Attention Network (GAT).[23] GATs use attention mechanisms to update pre-defined information about nodes depending on their connectivity. Any given attention mechanisms *a* is computed using the following equation from the original publication[23]:

$$a_{ij} = \frac{exp(e_{ij})}{\sum_{k \in N_i} exp(e_{ik})}$$

where i is a given cell/ligand/receptor node, *j/k* are nodes in the direct neighbourhood of node i, N_i is the total set of nodes and e_{ij}/e_{ik} are coefficients computed using input node embeddings and trainable weight matrices for each node. Attention mechanisms are then summed for all neighbouring nodes to obtain an updated node embedding for node *i* **(Suppleme Figure 2)**. Attention mechanisms are updated for 100 epochs during the training process. After being passed through a logistic sigmoid function (utilized for binary classification), node embeddings for ligand and receptor nodes are utilised to minimise cross entropy loss[62] against assigned class labels for input ligands and receptors (1 if a ligand/receptor participates in a validated intercell interaction, 0 if not). The loss function can be described as the following (Equation 2):

$$L_{CE} = -\sum_{i=1}^{n} t_i log(p_i)$$

Where L_{CE} is the cross-entropy loss, N is the number of nodes, t_i is the truth label (in this case, 1 or 0) and p_i is the sigmoid probability for that node from GraphComm.

Inference using trained GAT model

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 input 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.

Data Availability

The sequenced Stage 6 *Drosophila* embryos can be found via the GEO accession code GSE95025[63]. The sequenced embryonic mouse brain cells at Day E14.5 can be found via the GEO accession code GSE133079[29]. Pre and Post treatment data of PC9 cell lines treated with osimertinib can be found via the GEO accession code GSE150949[30]. Pre and Post treatment data of small cell lung cancer CTC derived xenografts can be found at the GEO accession code GSE138267[37]. Spatial transcriptomics data[47] 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 a Code Ocean capsule.[64]

Research Reproducibility

Full reproducibility of all results is available at the corresponding Code Ocean Capsule[64]. Model checkpoints, as well as results used to generate published figures are available for analysis. The Code Ocean capsule is configured to, at the click of a button, run pre-saved models on all inference datasets included in this manuscript.

Acknowledgements

We thank D. Forster, H. Cui, H. Maan and I. Smith for their insightful comments. Infographics in figures were created with Biorender. Experiments were run using resources from the Vector Institute.

Disclosures

SH reports funding from Novo Nordisk and Askbio GmbH, and is a co-founder of Sequantrix GmbH. BHK is a shareholder and paid consultant for Code Ocean Inc. ES is a paid consultant for Code Ocean Inc.

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. 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).
16. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2019).
17. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. Preprint at https://doi.org/10.1101/2020.07.21.214387.
18. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
19. 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).
20. Tsitsiridis, G. *et al.* CORUM: the comprehensive resource of mammalian protein complexes–2022. *Nucleic Acids Res.* **51**, D539–D545 (2022).
21. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. *arXiv [cs.SI]* Preprint at http://arxiv.org/abs/1607.00653 (2016).
22. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
23. Veličković, P. *et al.* Graph Attention Networks. *arXiv [stat.ML]* Preprint at http://arxiv.org/abs/1710.10903 (2018).
24. GeneCards Human Gene Database. PCNA Gene - GeneCards.

https://www.genecards.org/cgi-bin/carddisp.pl?gene=PCNA.

25. GeneCards Human Gene Database. LAMP1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=LAMP1&keywords=LAMP1.
26. Tang, F. & Christofori, G. The cross-talk between the Hippo signaling pathway and autophagy:implications on physiology and cancer. *Cell Cycle* **19**, 2563–2572 (2020).
27. GeneCards Human Gene Database. WNT4 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=WNT4&keywords=WNT4.
28. GeneCards Human Gene Database. PDE1C Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=PDE1C&keywords=PDE1C.
29. Sheikh, B. N. *et al.* Systematic Identification of Cell-Cell Communication Networks in the Developing Brain. *iScience* **21**, 273–287 (2019).
30. Oren, Y. *et al.* Cycling cancer persister cells arise from lineages with distinct programs. *Nature* **596**, 576–582 (2021).
31. 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).
32. AlMusawi, S., Ahmed, M. & Nateri, A. S. Understanding cell-cell communication and signaling in the colorectal cancer microenvironment. *Clin. Transl. Med.* **11**, e308 (2021).
33. Ramirez, M. *et al.* Diverse drug-resistance mechanisms can emerge from drug-tolerant cancer persister cells. *Nat. Commun.* **7**, 10690 (2016).
34. GeneCards Human Gene Database. TGFB1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=TGFB1&keywords=TGFB1.
35. Huang, A.-L. *et al.* TGF-β1 protein expression in non-small cell lung cancers is correlated with prognosis. *Asian Pac. J. Cancer Prev.* **15**, 8143–8147 (2014).
36. Wang, C. *et al.* Acquired Resistance to EGFR TKIs Mediated by TGFβ1/Integrin β3

Signaling in EGFR-Mutant Lung Cancer. *Mol. Cancer Ther.* **18**, 2357–2367 (2019).

37. Stewart, C. A. *et al.* Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. *Nat Cancer* **1**, 423–436 (2020).
38. GeneCards Human Gene Database. ANXA2 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=ANXA2&keywords=ANXA2.
39. Wang, C.-Y. *et al.* Annexin A2 silencing induces G2 arrest of non-small cell lung cancer cells through p53-dependent and -independent mechanisms. *J. Biol. Chem.* **287**, 32512–32524 (2012).
40. GeneCards Human Gene Database. YBX1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=YBX1&keywords=YBX1.
41. Johnson, T. G., Schelch, K., Mehta, S., Burgess, A. & Reid, G. Why Be One Protein When You Can Affect Many? The Multiple Roles of YB-1 in Lung Cancer and Mesothelioma. *Front Cell Dev Biol* **7**, 221 (2019).
42. GeneCards Human Gene Database. NRP1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=NRP1&keywords=NRP1.
43. Ding, Z. *et al.* Neuropilin 1 modulates TGF□β1□induced epithelial□mesenchymal transition in non□small cell lung cancer. *Int. J. Oncol.* **56**, 531–543 (2020).
44. GeneCards Human Gene Database. MCL1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MCL1&keywords=MCL1.
45. Michels, J. *et al.* MCL-1 dependency of cisplatin-resistant cancer cells. *Biochem. Pharmacol.* **92**, 55–61 (2014).
46. Zhang, F. *et al.* Simultaneous targeting of ATM and Mcl-1 increases cisplatin sensitivity of cisplatin-resistant non-small cell lung cancer. *Cancer Biol. Ther.* **18**, 606–615 (2017).
47. Kuppe, C. *et al.* Spatial multi-omic map of human myocardial infarction. *Nature* **608**, 766–777 (2022).
48. GeneCards Human Gene Database. ANXA1 Gene - GeneCards.

https://www.genecards.org/cgi-bin/carddisp.pl?gene=ANXA1&keywords=ANXA1.

49. GeneCards Human Gene Database. FPR3 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=FPR3&keywords=FPR3.
50. Qin, C. X. *et al.* Cardioprotective Actions of the Annexin-A1 N-Terminal Peptide, Ac2-26, Against Myocardial Infarction. *Front. Pharmacol.* **10**, 269 (2019).
51. GeneCards Human Gene Database. SLIT3 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=SLIT3&keywords=SLIT3.
52. GeneCards Human Gene Database. ROBO4 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=ROBO4&keywords=ROBO4.
53. Dai, C., Gong, Q., Cheng, Y. & Su, G. Regulatory mechanisms of Robo4 and their effects on angiogenesis. *Biosci. Rep.* **39**, (2019).
54. Paul, J. D. *et al.* SLIT3-ROBO4 activation promotes vascular network formation in human engineered tissue and angiogenesis in vivo. *J. Mol. Cell. Cardiol.* **64**, 124–131 (2013).
55. Wu, X., Reboll, M. R., Korf-Klingebiel, M. & Wollert, K. C. Angiogenesis after acute myocardial infarction. *Cardiovasc. Res.* **117**, 1257–1273 (2021).
56. Dimitrov, D. *et al.* LIANA+: an all-in-one cell-cell communication framework. *bioRxiv* 2023.08.19.553863 (2023) doi:10.1101/2023.08.19.553863.
57. GeneCards Human Gene Database. HP Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=HP&keywords=HP.
58. GeneCards Human Gene Database. CD163 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CD163&keywords=CD163.
59. Zhang, J., Ding, W., Liu, J., Wan, J. & Wang, M. Scavenger Receptors in Myocardial Infarction and Ischemia/Reperfusion Injury: The Potential for Disease Evaluation and Therapy. *J. Am. Heart Assoc.* **12**, e027862 (2023).
60. Omnipath. *PyPI* https://pypi.org/project/omnipath/.
61. Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. in

ICLR Workshop on Representation Learning on Graphs and Manifolds (2019).

62. 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).
63. Karaiskos, N. *et al.* The Drosophila embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
64. Code Ocean. https://codeocean.com/capsule/8269062/tree/v4

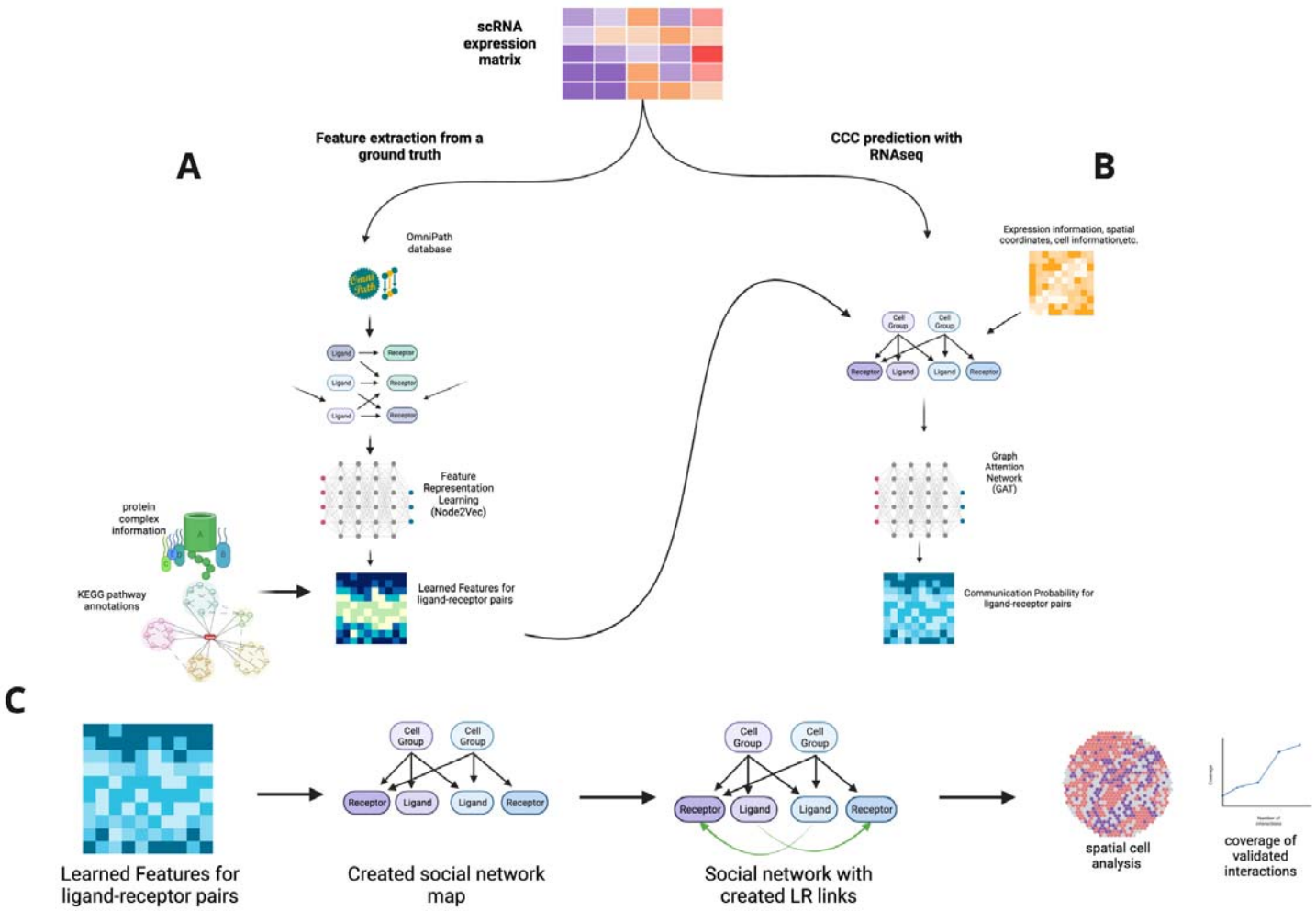


Figure 1: Schematic outlining the architecture of GraphComm to make CCC predictions from scRNAseq data **(A)** GraphComm utilises a scRNAseq dataset with a curated ligand-receptor database to construct a directed graph reflective of the CCC ground truth. Feature Representation is implemented to extract positional information for ligands and receptors within the directed graph and scaled accordingly with protein complex and pathway information **(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.

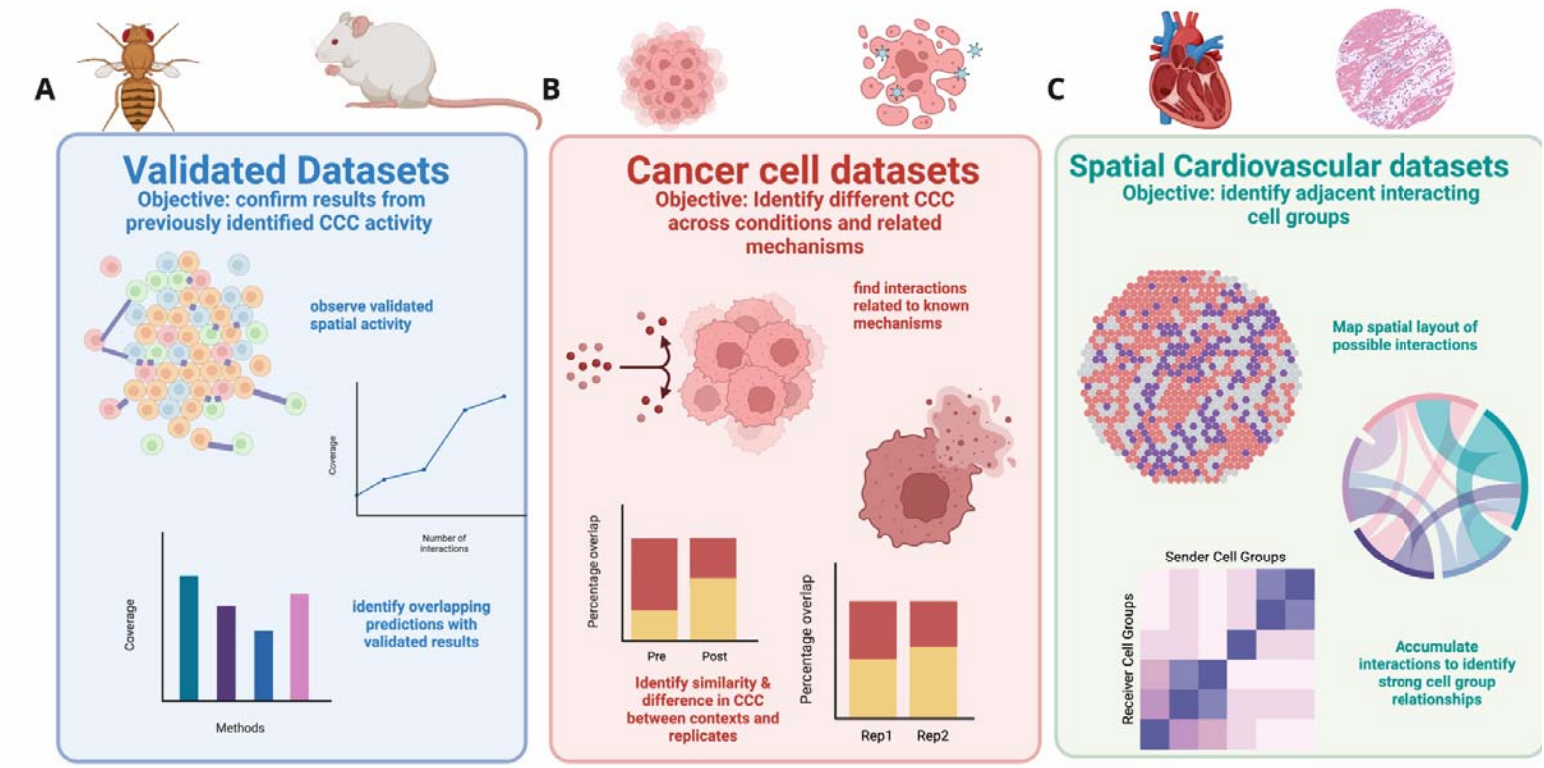


Figure 2: Schematic of the validated datasets that GraphComm performs inference on and the purpose for their respective benchmarks. **(A)** Datasets with validated CCC activity, from stage 6 *Drosophila* embryo and embryonic mouse brain, were used in GraphComm prediction to confirm predictions aligned with the previously validated results. Measures of success in those benchmarks include the plotting of spatial activity and top-ranked interactions to compare with the ground truth.**(B)** Datasets of treated cancer cells, from PC9 cell lines and small cell lung cancer xenografts, were used in GraphComm prediction to identify how our framework can identify different CCC across conditions and connect communication activity to related, well-studied mechanisms. Benchmarks will compare the similarity and difference in interaction overlap between datasets of different conditions (pre and post treatment) and biological replicates. Furthermore, validity in predicted interactions can be strengthened by highlighting ligands and receptors that have been previously studied in cancer cell drug resistance and sensitivity mechanisms. **(C)** Datasets containing spatial transcriptomics, from human hearts following myocardial infarction, were used in GraphComm prediction to identify potentially interacting adjacent cell groups. With location information within an individual slide, the likelihood of an interaction can be correlated to its predicted communication probability. By also identifying the number of interactions in one zone across all available slides, crosstalk between two cell groups can be analysed for their relevance towards the dataset.

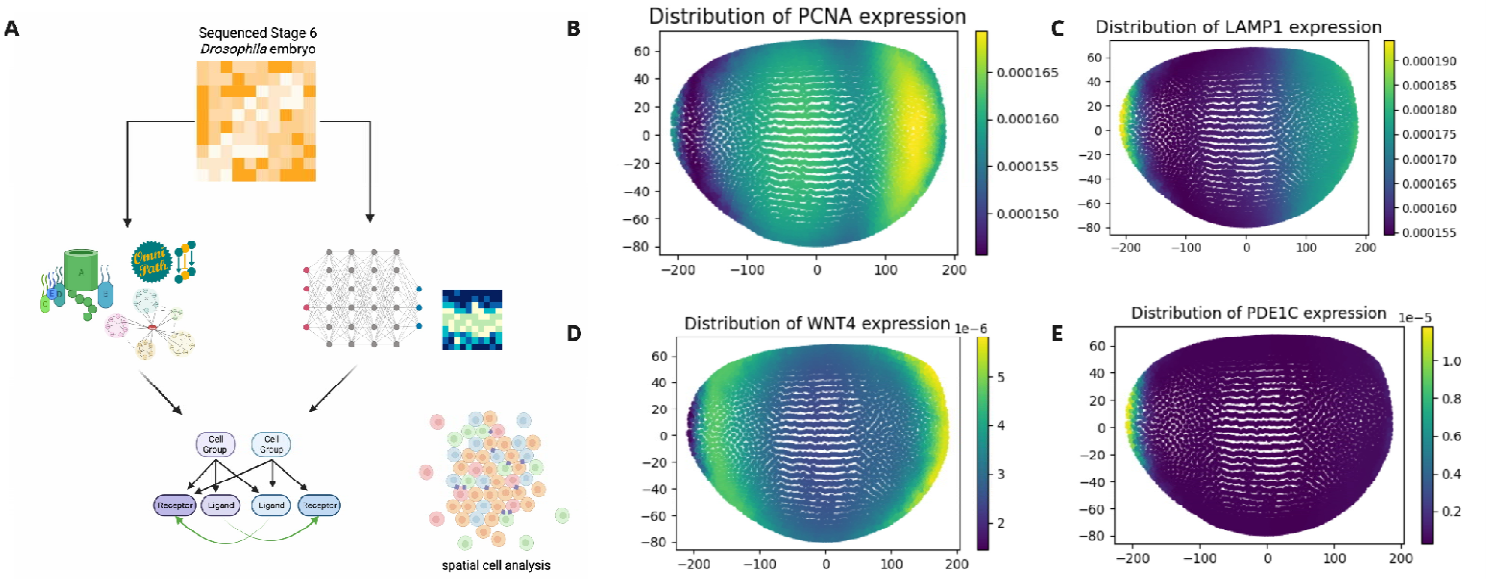


Figure 3: (A) GraphComm architecture can identify previously validated pathway activity from in-situ hybridization experiments. With a scRNA dataset of stage 6 *Drosophila* embryo as input, communication probabilities can be used to identify top-ranked interactions of ligands in the dataset for subsequent spatial plotting analysis. **(B,C)** The spatial gene expression of a top ranked interaction between the ligand PCNA and the receptor LAMP1 **(D,E)** The spatial gene expression of a lower ranked interaction between the ligand WNT4 and PDE1C.

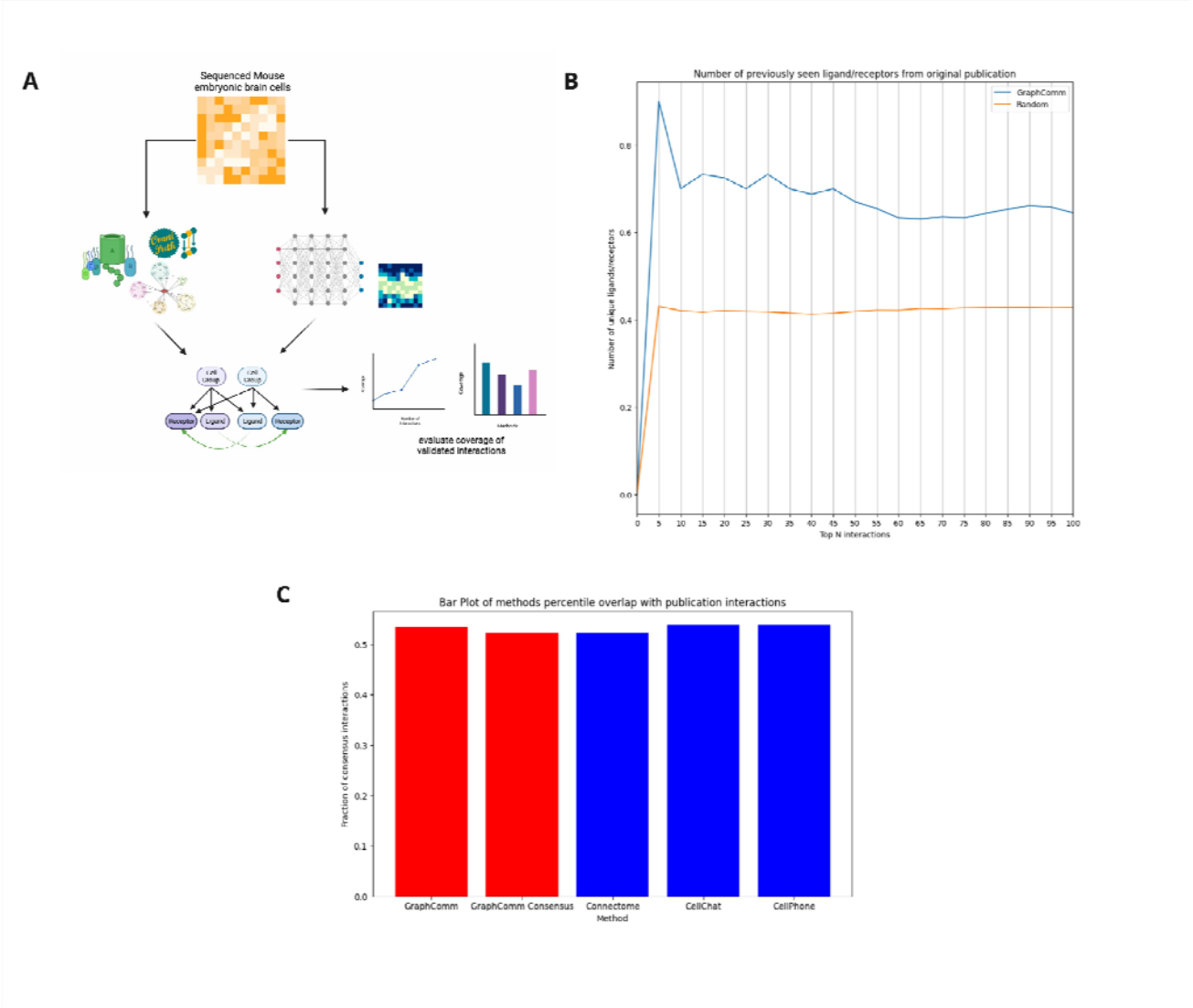


Figure 4: (A) GraphComm can identify interactions previously validated *in silico* in an embryonic mouse brain **(B)** In comparison to 100 randomised iterations, GraphComm can identify more interactions containing previously validated ligands and receptor in the top 100 interactions with ligand and receptor candidates **(C)** In the overall set the performance of predicting previously validated interactions with GraphComm is comparable to other methods, with GraphComm trained on both intracell and intercell datasets performing better than GraphComm trained on just the consensus dataset.

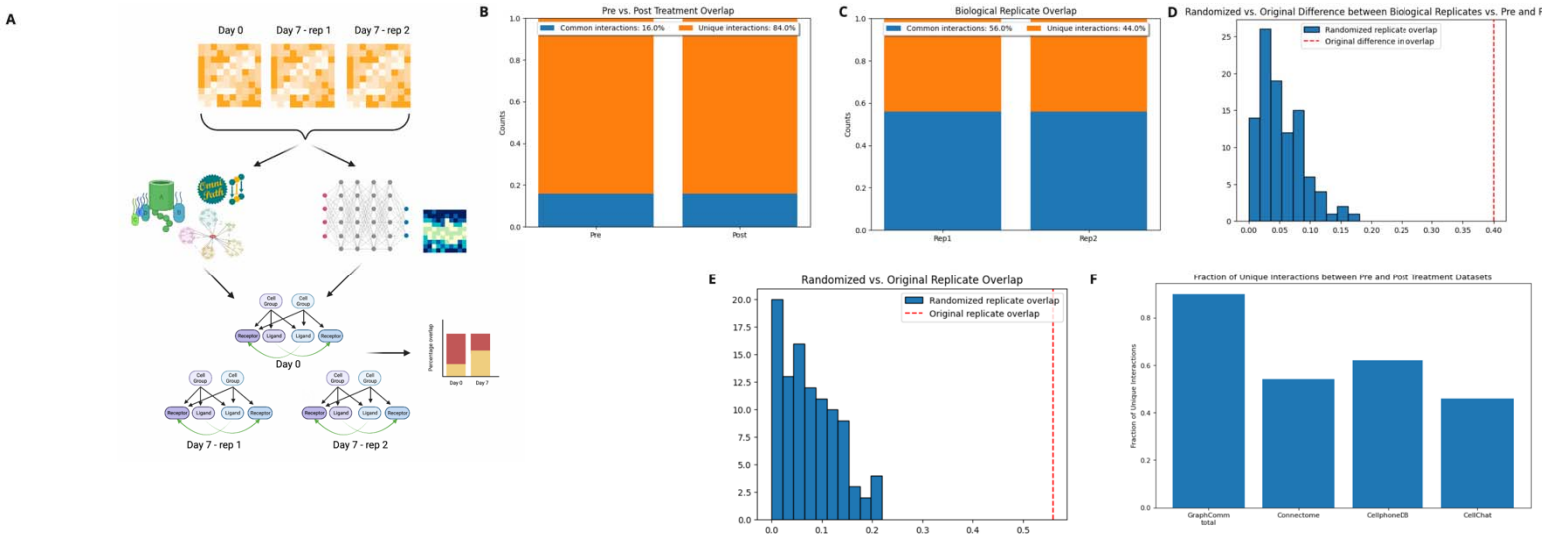


Figure 5: **(A)** GraphComm can identify differences in interactions among datasets that have the potential to be indicative of change in condition. Using multiple datasets of PC9 lung adenocarcinoma cell lines sequenced at different timepoints before and after treatment with osimertinib including two 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 16% **(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 56% **(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.

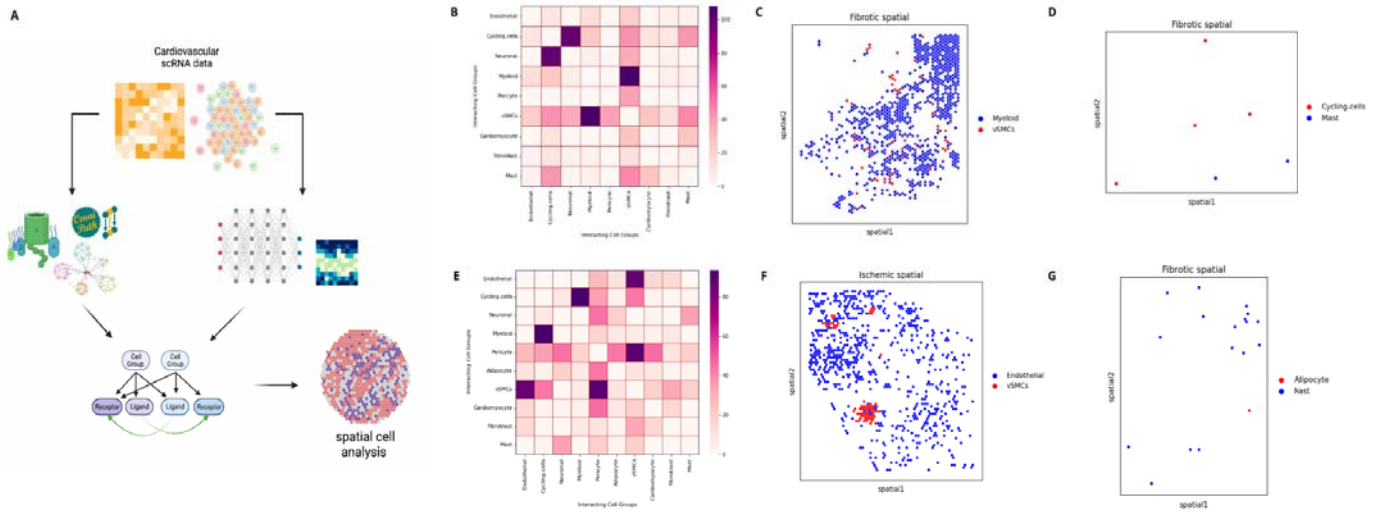


Figure 6: **(A)** GraphComm architecture can predict CCC occurring in spatially adjacent cell groups. Utilising spatial transcriptomics data from the fibrotic region of a patient sample, GraphComm can identify ligand-receptor interactions as well as identify potential source and destination cell groups. Then, the nodes of interest can be used to visualise two interaction cell groups and their respective proximity. **(B)** heatmap displaying number of interactions between any two cell groups, across all fibrotic slides **(C)** Spatial map only displaying cells in top interacting cell groups found by GraphComm in fibrotic slides, Myeloid and vSMC 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 interacting cell groups found by GraphComm in fibrotic slides, Cycling cells and Mast cells. **(E)** Spatial heatmap displaying number of interactions between any two cell groups, across all ischemic slides **(F)** Spatial map only displaying cells in top interacting cell groups found by GraphComm in ischemic slides, Endothelial and vSMC cells. From this spatial plot alone, the adjacency of these two cell groups and the potential for interactions can already be observed **(G)** Spatial map only displaying cells in low interacting cell groups found by GraphComm in ischemic slides, Adipocyte cells and Mast cells

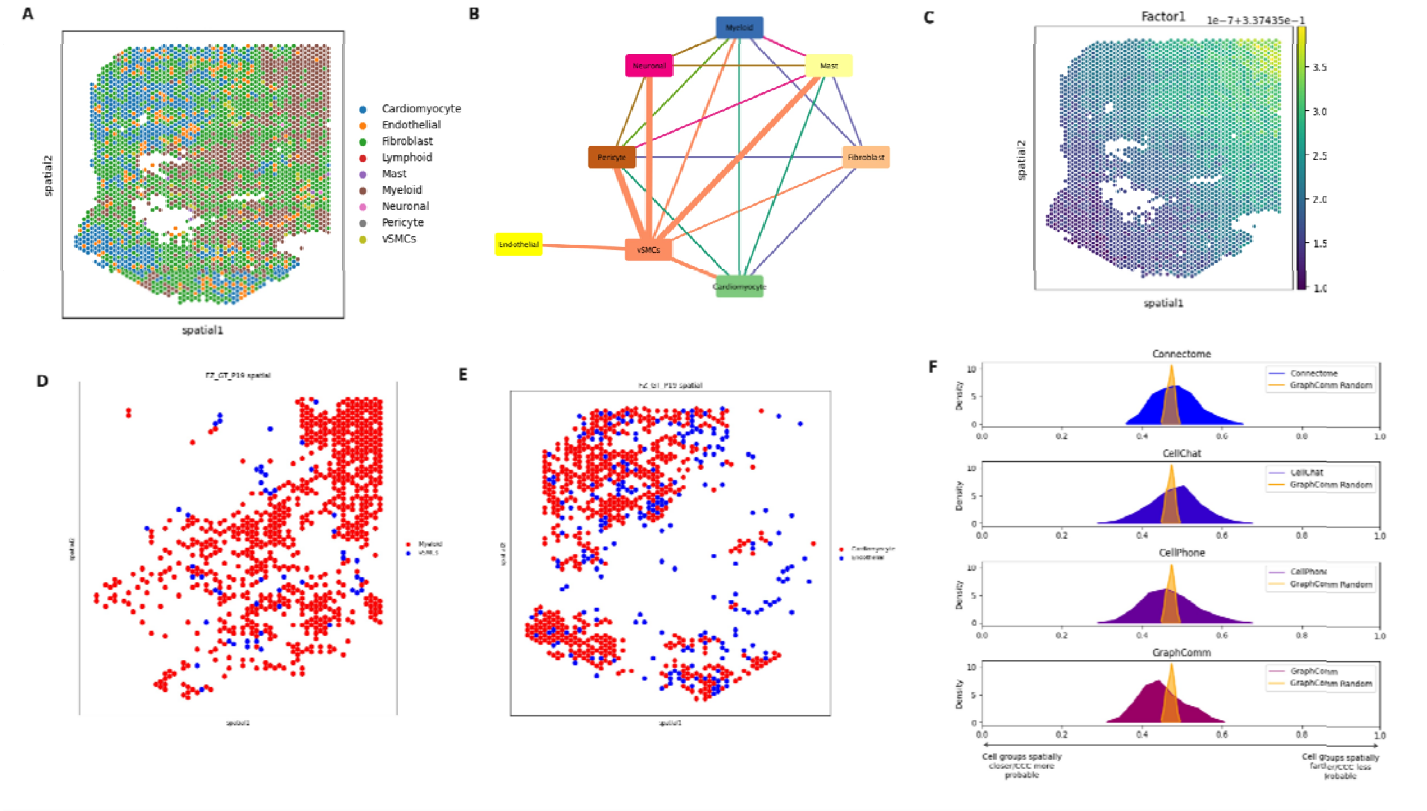
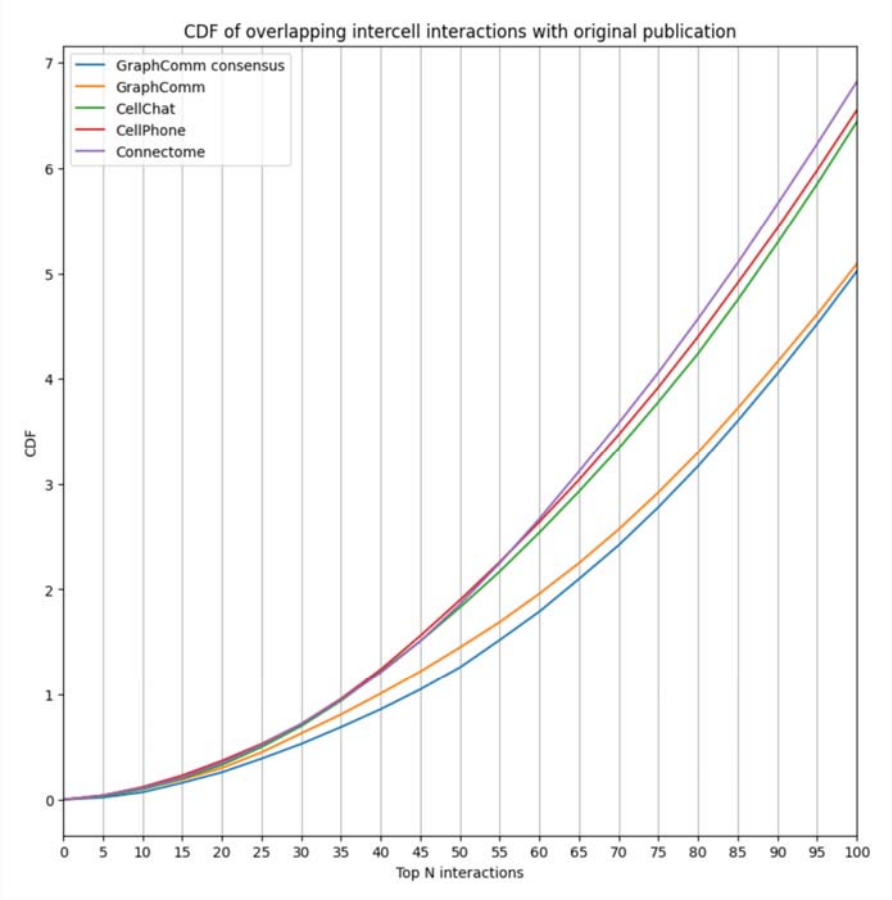
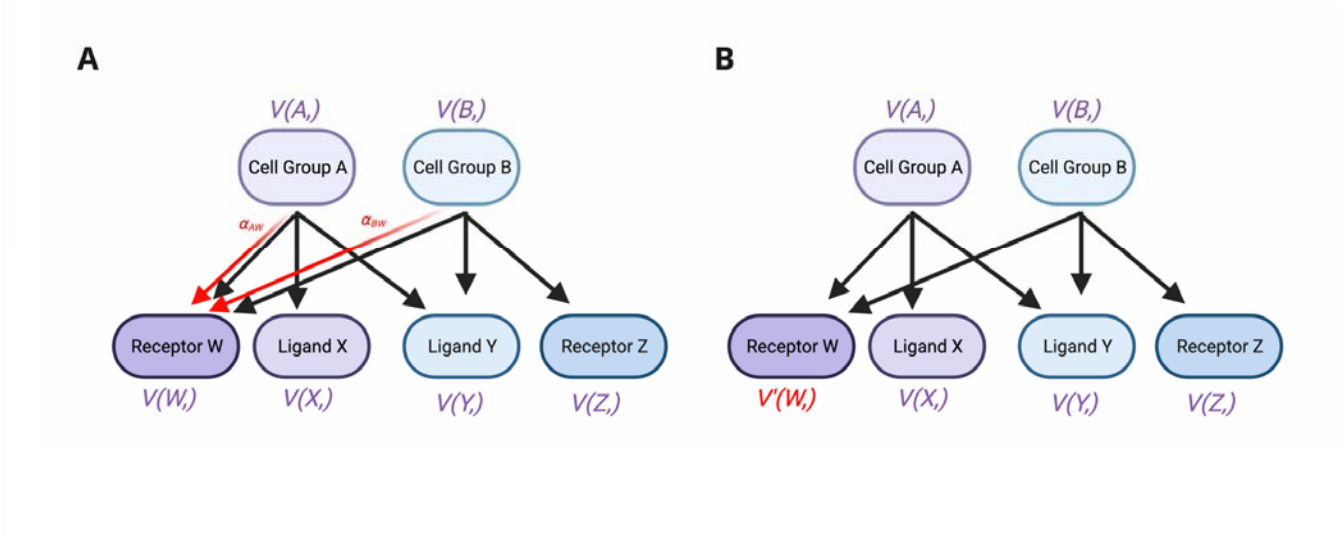


Figure 7: (A) Spatial coordinates of all cells in the dataset and colour coded by the cell-type in the *FZ_GT_P19* slide. **(B)** Weighted circle graph demonstrating the most prevalent cell group interactions found in the top 100 interactions predicted by GraphComm **(C)** Results from NMF done by LIANA+, detailing regions of the fibrotic slide most indicative of cell-cell communication **(D)** Spatial map only displaying cells in top interacting cell groups found by GraphComm, vSMC and Myeloid cells. From this spatial plot alone, the adjacency of these two cell groups and the potential for interactions can already be observed **(E)** Spatial map only displaying cells in low-ranking interacting cell groups found by GraphComm, Endothelial and Cardiomyocyte cells. From this spatial plot alone, the adjacency of these two cell groups and the potential for interactions can already be observed **(F)** 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.



Supplementary Figure 1:

In the top 100 interactions, GraphCom predicts slightly less interactions that overlap with the publication in comparison to other methods. Potential reasoning behind this result includes GraphComm's prioritisation of previous intercellular interactions that were not originally investigated



Supplementary Figure 2:
The mechanisms behind the GAT architecture, demonstrating how embedding information is updated for nodes during model training. **(A)** Input graph during GAT training, during embedding update of a single node, Receptor W (all nodes are updated during one training epoch). A given node *N* is annotated with V(*N,)*, representing the full vector of embeddings (Node *N's* numerical relation to all other nodes in the graph). From each first order neighbouring node of Receptor W, the attention mechanism with respect to Receptor W is calculated and concatenated to the existing Receptor W embedding. **(B)** Input graph post update of Receptor W embedding, *V(W,)* has been updated to *V'(W,)* to reflect the updated embedding. The process is continued for each node in the graph, which completes the epoch. Both target nodes are updated from their source nodes and vice versa (i.e. embeddings from ligands and receptors are used to update cell group embeddings).

Tables

Type of dataset	Organism	Tissue type	Modality	Number of cells/spots	Number of Major cell-types	Reference
Previously validated CCC activity	*Drosophila melanogaster*	Stage 6 embryo	Single-cell transcriptomics	5835	N/A	Karaiskos et al. 2017[63]
	Mus musculus	Embryonic brain	Single-cell transcriptomics	2304	N/A	Sheikh et. al 2019[29]
Not validated for CCC activity	*Homo sapiens*	Heart Left Ventricle	Spatial transcriptomics	FZ_GT_P19: 3100 FZ_GT_P4: 4253 FZ_P14:3175 FZ_P18:2551 FZ_P20:2410 RZ_FZ_P5:3082 GT_IZ_P13:1890 GT_IZ_P15:3572 GT_IZ_P9:4361 IZ_BZ_P2:4203 IZ_P10:3646 IZ_P15:3083 IZ_P16:2713 IZ_P3:3771	FZ_GT_P19:9 FZ_GT_P4:11 FZ_P14:9 FZ_P18:9 FZ_P20:11 RZ_FZ_P5:7 GT_IZ_P13:7 GT_IZ_P15:9 GT_IZ_P9:10 IZ_BZ_P2:11 IZ_P10:9 IZ_P15:9 IZ_P16:9 IZ_P3:11	Kuppe et al. 2022[47]
		PC9 lung cell lines	Single-cell transcriptomics	Day 0: 6778 Day 7 rep1:4770 Day 7 rep2:3515	N/A	Oren et al. 2021[30]
		Small cell lung cancer xenografts (treatment naive and resistant)	Single-cell transcriptomics	Treatment Naive pre-treatment: 40166 Treatment Resistant pre-treatment: 4684 Treatment Naive post-treatment:2176 Treatment Resistant post-treatment:1050	N/A	Stewart et. al 2020[37]

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

Supplementary Information

Supplementary Tables 1-4: The top 100 validated intercell interactions from the pre and post treatment small cell lung cancer xenografts, both treatment-naive and treatment resistant models.