# CCIDeconv: Hierarchical model for deconvolution of subcellular cell-cell interactions in single-cell data

Rojashree Jayakumar,[1,2,3] Pratibha Panwar,[1,2,3] Jean Yee Hwa Yang[1,2,3] and Shila Ghazanfar[1,2,3] *

[1]School of Mathematics and Statistics, Faculty of Science, The University of Sydney, Camperdown, New South Wales, 2006, Australia
[2]Sydney Precision Data Science Centre, The University of Sydney, Camperdown, New South Wales, 2006, Australia
[3]Charles Perkins Centre, The University of Sydney, Camperdown, New South Wales, 2006, Australia

*Corresponding author: shila.ghazanfar@sydney.edu.au

## Abstract

**Motivation: Cell-cell interaction (CCI) underlies several fundamental mechanisms including development, homeostasis and disease progression. CCI are known to be localised to specific subcellular regions, for example, within the cytoplasms of cells. With the emergence of subcellular spatial transcriptomics technologies (sST), there is an opportunity to attribute CCI to subcellular regions. We aimed to deconvolute CCI to subcellular CCI (sCCI) in non-spatial single cell transcriptomics data (i.e. scRNA-seq) datasets using a modified CCI score from CellChat.**
**Results: By calculating the sCCI score specific to cytoplasm and nucleus in nine publicly available sST datasets, we identified unique nucleus-nucleus and cytoplasm-cytoplasm sCCI. Then, we deconvolved the communication score to subcellular regions by using a hierarchical classification and regression model which we name as CCIDeconv. We performed leave-one-dataset-out cross-validation across nine datasets over a range of different tissue types from human samples. We observed that training across many different tissue types resulted in robust deconvolution performance in an unseen dataset. As the number of training datasets increased, models trained without spatial features achieved similar performance as models including spatial features. This implied the potential for accurate prediction of sCCI events from even scRNA-seq with large numbers of training datasets. Overall, we offer a method towards attributing CCI events to subcellular regions. This method can allow researchers in dissecting sCCI patterns to gain insights in underlying biology in a range of tissues covering health and disease.**

**Keywords:** Cell-Cell Communication, Single Cell RNA-seq, Predictive Modeling, Bioinformatics

## 1. Introduction

CCI is important in understanding biological systems and occurs through intracellular and intercellular mechanisms. It involves signal generation, transmission, and subsequently signal transduction (Su et al., 2024). These events are mediated by ligands on the sending cell and receptors on the receiving cell. Ligand and receptors (LR) can range from extracellular proteins to secreted molecules such as ions, cytokines and hormones. Therefore, cells communicate with each other using a diverse range of molecules and signal various downstream pathways that result in altered gene expression (Armingol et al., 2020).
With a wealth of high-throughput single-cell transcriptomics data, there have been tremendous efforts in the inference of CCI at the individual cell and cell type levels. Given a database of putative LR pairs and complexes, e.g. OmniPath (Türei et al., 2016), CellPhoneDB (Troulé et al., 2025), and CellChatDB (Jin et al., 2021), and a single-cell dataset, methods such as CellChat, NicheNet (Browaeys et al., 2020), and CellPhoneDB identify CCI that are present in the given tissue. Methods such as MISTy (Tanevski et al., 2022) and NicheCompass (Birk et al., 2025) have been developed to utilize the additional spatial coordinates of cells to perform inference of CCI events. Similarly, methods originally developed for single-cell dissociated data have undergone additional development to account for spatial coordinates, including CellChat (Jin et al., 2025) and

© The Author(s) . Copyright and licence statements to be updated by the publisher during production.

CellPhoneDB (v5). Some methods such as DeepLinc (Li and Yang, 2022) identify de novo CCI without the need of a curated database. Methods like CellChat calculate the communication score as a function of the expression and assess significance by permutation testing, while other methods like NicheNet use network based approaches to calculate the scores.
Interestingly, such approaches have not attempted to capitalise on the potential additional resolution afforded by some spatial transcriptomics technologies. sST are technologies that enable quantification of gene expression at the subcellular region level. While there are sequencing-based spatial transcriptomics platforms that offer subcellular resolution through DNA nanoballs (Stereo-seq (Zhao et al., 2025)) or extremely high resolution chips (10X Genomics Visium HD (Monica Nagendran, 2023), sST are primarily obtained through the aforementioned imaging platforms.
Subcellular localisation of LR can determine where CCI is initiated, which signalling pathways are activated, influence disease mechanisms, and guide therapeutic targeting. Literature and experimental evidence show that certain interactions occur in the cytoplasm, like the interactions between tumor-stromal communication (Gardiner and Cukierman, 2022). Some occur on the surface; for example, microglial activation is based on LR recognition on the cell surface in Alzheimer's disease (Chen et al., 2006). G-protein-coupled receptors (GPCRs) can initiate signaling not only at the plasma membrane but also from intracellular organelles, and these distinct locations can produce different downstream effects and cellular functions (Bock et al., 2025). Vargas et al. (2023) showed that engaging serotonin 2A receptors inside neurons promotes growth of new connections but engaging the same receptor on the surface of nerve cells does not. Additionally, mechanisms such as extracellular vesicles mediated delivery of ligands inside the cell (Lorico et al., 2025) further support the existence of communication at subcellular levels.
Here, we show that CCI differs between the subcellular regions and present CCIDeconv. CCIDeconv effectively deconvolutes CCI scores into interactions primarily occurring in nucleus or cytoplasm leveraging information from sST. CCIDeconv provides insight into the location of CCI, and therefore can be used to further refine their communication scores.

## 2. Methods

### 2.1. Calculation of CellChat based communication score

The CCIDeconv communication score is a modified score derived from CellChat. LR pairs from the CellChatDB.human database are used to calculate the scores. We modified the CellChat formula by setting the components of the CellChat formula that models agonists and antagonists expression to a value of 1.

#### 2.1.1. Communication score with spatial information

The communication score is calculated as

$$C_{cell} = \frac{L_i R_j}{K_h + L_i R_j} S_{i,j}$$

where $L_i$ is the expression level of a ligand in cell type $i$, $R_j$ is the expression level of a receptor in cell type $j$, $S_{i,j}$ is the spatial distance between cell types $i$ and $j$ and $K$ is the dissociation coefficient and $h$ is the Hill's constant (Gesztelyi et al., 2012). $R_j$ is weighted by expression of co-stimulatory ($RA_j$) and co-inhibitory receptors ($RI_j$) in cell type $j$ and is calculated as $R_j = \frac{1+RA_j}{1+RI_j}$. $S_{i,j}$ is calculated using the computeRegionDistance() function from CellChat. We set $K = 0.5$ and $h = 1$ for all our scores. Additionally we define subcellular specific CCI scores for the cytoplasm-cytoplasm and nucleus-nucleus communication:

$$C_{cyt} = \frac{L_i^{cyt} R_j^{cyt}}{K_h + L_i^{cyt} R_j^{cyt}} S_{i,j}$$

$$C_{nuc} = \frac{L_i^{nuc} R_j^{nuc}}{K_h + L_i^{nuc} R_j^{nuc}} S_{i,j}$$

where $L_i^{cyt}$, $L_i^{nuc}$, $R_j^{cyt}$ and $R_j^{nuc}$ is the ligand and receptor expression level in the corresponding subcellular regions.

#### 2.1.2. Communication score without spatial information

For deconvolution of single cell data, we effectively set the spatial distance $S_{i,j}$ to 1 when calculating the communication score. It is calculated as

$$C_{cell} = \frac{L_i R_j}{K_h + L_i R_j}.$$

### 2.2. Overview of CCIDeconv

#### 2.2.1. CCIDeconv architecture

The architecture is a hierarchical supervised machine learning framework with a classification followed by two regression models. A classifier classifies each CCI event detected in the whole cell into a terminal class $\boldsymbol{o}$ and a separable class $\boldsymbol{x}$. Class $\boldsymbol{o}$ represents CCI that have low expression signal in subcellular regions that cannot be deconvoluted. Class $\boldsymbol{x}$ represents CCI whose communication scores can be deconvoluted. CCI in class $\boldsymbol{x}$ undergo two rounds of regression to obtain two communication scores, attributing to the nucleus and cytoplasm communication. Two different procedures using the same

CCIDeconv architecture is implemented in this paper:
(i) Spatial procedure (SP): $S_{i,j}$ is included in the communication score and is a feature along with the other features mentioned in 2.2.2.
(ii) Single-cell procedure (ScP): This procedure does not include $S_{i,j}$ both in the communication score and as a feature.

#### 2.2.2. Feature scaling and encoding

The input features to the model are communication scores in cell, Hill function of the LR expression, receiver and sender cell types, ligand and receptor HGNC symbols as well as subcellular location and molecular information of ligands from CellChatDB and Human Protein Atlas (HPA) (The Human Protein Atlas; Uhlén et al., 2015). We scale the continuous features to zero mean and unit variance using StandardScaler from the scikit-learn library (Gramfort et al., 2011) in Python (version 3.10). All other categorical features are target encoded (Pargent et al., 2022). The location features are transformed into a binary indicator matrix using the MultiLabelBinarizer. The encoding of location features is based on the model selection procedure which we have detailed in section 2.4.2.

#### 2.2.3. Model specification

The CCIDeconv model uses a voting classifier between Random Forest and XGBoost and XGBoost regressor. The choice of the model is according to the evaluation of the candidate approaches detailed in section 2.4.

#### 2.2.4. Hyperparameter tuning

For all models we optimize hyperparameters using a Bayesian optimization method (Garnett) implemented via the BayesianOptimization package. The parameters used for tuning and the search space are detailed in the Supplementary Methods 1.1.

### 2.3. Preprocessing datasets

We obtained nine sST datasets from the 10X Xenium platform (Supplementary Methods 1.2.1) and they were imported into R using the readXenium() function from the MoleculeExperiment package (Peters Couto et al., 2023). We computed SpatialExperiment objects for each subcellular region by aggregating the data using countMolecules(), with boundariesAssay = "cell" to obtain cell-level counts and boundariesAssay = "nucleus" to obtain nuclear counts. Cytoplasmic counts were calculated as the difference between the counts in the cell and nucleus. Therefore we obtained three SpatialExperiment objects (cell, cytoplasm and nucleus) for each dataset. Each SpatialExperiment object underwent the following processing steps. We used addPerCellQCMetrics() function from the scuttle package (McCarthy et al., 2016) to compute quality control metrics. Cells with total counts $\leq 20$ and expressed genes $\leq 10$ were removed. Negative-control probes were excluded prior to downstream analysis. Cells were normalised using logNormCounts() function from the scuttle package. We did cell type annotation with corresponding reference datasets (detailed in Supplementary Methods 1.2.2) using the singleR() function from singleR package (Aran et al., 2019). We then followed the spatial data processing from CellChat. The communication score was calculated as described in 2.1. As a result we obtained significant CCI (P-value $\leq 0.05$) for each dataset across the cell, cytoplasm and nucleus.

### 2.4. Evaluation of candidate models for CCIDeconv

#### 2.4.1. Leave one group out cross-validation

We evaluated CCIDeconv using Leave one group out cross-validation (LOGO-CV) (Gramfort et al., 2011) on the nine sST datasets. We used the SP procedure to assess performance on the sST dataset and the SCP procedure to assess if CCIDeconv can be applied to single cell datasets. We conducted scaling and target encoding steps within training data of each fold to prevent information leakage.

#### 2.4.2. Model Selection from different baselearner configurations and encoding location features

We evaluated seven different models with different base learning configurations and feature encoding strategies of the subcellular location features from CellChatDB and Human Protein Atlas. To select the best model we used LOGO-CV. The details of the various model specifications are given in Supplementary Table 1.

#### 2.4.3. Evaluation metrics

We evaluated the model performance with a series of metrics such as accuracy, Area Under the Curve (AUC), Recall (macro), $R^2$ and Normalized Root Mean Squared Error (NRMSE) calculated on unknown test datasets. NRMSE was calculated by normalising the mean squared error by the standard deviation of the response (communication scores in nucleus or cytoplasm).

**Table 1** Evaluation metrics of the CCIDeconv (LOGO-CV) across the nine datasets using SP procedure.

| Metric | Mean | Median | Std Dev | Std Err | IQR |
|---|---|---|---|---|---|
| AUC | 0.79 | 0.80 | 0.11 | 0.04 | 0.12 |
| Recall (macro) | 0.69 | 0.67 | 0.09 | 0.03 | 0.09 |
| $R^2$ cytoplasm | 0.87 | 0.88 | 0.07 | 0.02 | 0.05 |
| $R^2$ nucleus | 0.8 | 0.80 | 0.14 | 0.04 | 0.13 |
| NRMSE cytoplasm | 0.34 | 0.35 | 0.12 | 0.04 | 0.08 |
| NRMSE nucleus | 0.41 | 0.44 | 0.17 | 0.06 | 0.16 |
| Composite metric | 0.73 | 0.75 | 0.06 | 0.02 | 0.1 |

The composite metric is calculated as the geometric mean of the above metrics to assess performance.

$$\text{Composite metric} = \left( \text{AUC} \times \text{Macro Recall} \times \frac{R^2_{\text{cyt}} + R^2_{\text{nuc}}}{2} \times \frac{(1 - \text{NRMSE}_{\text{cyt}}) + (1 - \text{NRMSE}_{\text{nuc}})}{2} \right)^{1/4}$$

Where $\text{NRMSE}_{\text{cyt}}$ $\text{NRMSE}_{\text{nuc}}$ $R^2_{\text{cyt}}$ and $R^2_{\text{nuc}}$ represent the NRMSE and $R^2$ values of cytoplasmic and nuclear regression.

#### 2.4.4. Robustness analysis

To explore the robustness of CCIDeconv with respect to the number of datasets, we trained on every possible combination of the eight datasets with the ninth dataset as the test dataset. This resulted in 255 training combinations for each dataset. Scaling and target encoding steps are conducted within each combination to prevent information leakage. We used both SP and ScP procedures for each combination and assessed performance using the composite metric.

### 2.5. Application to scRNA-seq data

We applied CCIDeconv to lung cancer single cell data (Supplementary Methods 1.2.3.) by training on seven sST datasets using the ScP procedure. The two remaining sST datasets (Brain (AD) and Lung) were removed due to the region specific communication being primarily driven by the spatial features. We deconvoluted the cell CCI to obtain cytoplasm and nucleus specific CCI predictions. We performed a direct comparison of cytoplasm and nucleus CCI by calculating the log difference.

## 3. Results

### 3.1. Interactions between ligands and receptors vary between the subcellular regions

We hypothesised CCI varies across subcellular regions (Figure 1A). Using LR annotations from CellChatDB, we computed communication scores in cytoplasm and nucleus across nine sST datasets covering several tissue types. We observed marked differences in both the number of communicating LR pairs (Figure 1B, Supplementary Figure S1) and their communication scores in all datasets (Figures 1C, Supplementary Figures S2, S3). We identified 3,273 significant ($P$-value $< 0.05$) communicating LR pairs across the nine datasets. These LR pairs may recur, but are unique when defined by their cell-type to cell-type interactions. We identified 165 common pairs in cell and nucleus-nucleus communication and 568 common pairs in cell and cytoplasm-cytoplasm communication, indicating differences in sCCI. We also observed differences in communication score between cytoplasm to nucleus and nucleus to cytoplasm (Supplementary Figure S3).

We wondered if different components of CellChat communication score capture distinct, region-specific patterns of CCI. Principal component analysis (PCA) on various components of the CellChat communication score (i.e. Hill function, Spatial component, Agonist and Antagonist component) separated LR pairs based on the region where they were detected (Figure 1D). We found region specific communication patterns are driven by the Hill function of LR expression (Supplementary Figure S4).

Next, we assessed if the significant LRs in each subcellular region are enriched for pathways associated with CCI. We discovered that most enriched pathways were anchoring junctions and membrane pathways reflecting the function of cell communication (Supplementary Figure S5). Together with the observed differences in the number of communicating LR pairs, their communication scores, and PCA-based separation, these results demonstrate that CCI varies between cytoplasm and nucleus.

### 3.2. CCIDeconv's hierarchical modeling enables robust subcellular deconvolution of spatial communication scores

We identified distinct sCCI which motivated us to investigate whether we can deconvolute cell CCI to sCCI within cellular spatial profiles. Through CCIDeconv, we deconvolved the cell level communication scores to nucleus and cytoplasmic scores (Figure 2A). We evaluated the CCIDeconv across nine datasets using LOGO-CV through the SP procedure to account for spatial information. CCIDeconv is robust with a median composite metric of 0.75 across nine datasets (Figure 2B-D, Supplementary Figure S6). It achieved a mean AUC of 0.79 and macro recall of 0.69, with good performance in both the cytoplasm ($R^2 = 0.87$, NRMSE = 0.34) and the nucleus ($R^2 = 0.8$, NRMSE = 0.41) (Table 1). We found that CCIDeconv is computationally efficient as it takes less than 10 minutes to train and test the entire model on one dataset on a standard laptop. Thus, CCIDeconv is able to deconvolute the communication scores and attribute them to the nucleus and cytoplasm.

### 3.3. CCIDeconv is robust and stable across tissue types

We performed robustness analysis on CCIDeconv using the SP procedure to assess the stability of CCIDeconv across different tissue types (Methods 2.4.4.). We observed that 67.3% (ranging between 45.6% - 73.3% for each dataset) of LR pairs were classified correctly across the different training combinations (Figure 3A-D) with a median AUC of 0.80 (IQR: (0.66, 0.91), Figure 3E). The Brain (GL), Ovarian, and Pancreas datasets had a lower proportion of correctly classified LR pairs. Hence, we wondered if there were certain pairs that were classified incorrectly. In all the training combinations, 20% of the LR pairs were correctly classified (Figure 3C). Less than 5% of the LR pairs were incorrectly classified across all training combinations. Most of these belonged to the datasets mentioned above (Supplementary Figure S7A,B). LR pairs were grouped as "Easy" if they were correctly classified in at least 75% of combinations across all datasets, "Hard" if they were classified correctly in less than 25% of combinations or "Intermediate" otherwise. We found that 214 LR pairs were Easy, 24 were Hard, and 62 were Intermediate to classify (Figure 3F). Out of these, 153 LR pairs were classified correctly across all the training combinations and only 3 were always classified incorrectly. The APP_CD74 LR pair was detected 21 times irrespective of cell and tissue type and was always classified incorrectly (Figure 3G). The cytoplasmic $R^2$ was high with a low NRMSE (Figure 3F, Supplementary Table 2). Although the NRMSE-nucleus was higher than expected (median = 0.57), the $R^2$-nucleus was considerably higher (median = 0.67) indicating that CCIDeconv achieves good regression performance across different tissue types. Overall, the median composite metric was 0.66 and all other performance metrics were consistent across all the datasets (Supplementary Figure S7C,D). As a result, the model is highly stable and robust across tissue types.

### 3.4. Spatial information is important for performance when there are fewer training datasets

We then wished to determine the importance of spatial features in deconvoluting the communication score. We performed robustness analysis using both the SP and ScP procedures (Methods 2.4.4.). We observed that the composite metric was lower in models trained without the spatial features when the number of training datasets was less than 4 (Figure 3H). However as the number of training datasets increased, models trained with or without spatial features achieved similar results. We observed a similar pattern in terms of AUC and $R^2$ in cytoplasm (Supplementary Figure S8). In contrast we observed an opposite pattern in terms of macro recall and $R^2$ in the nucleus. Nevertheless, NRMSE was high in models trained without the spatial features when there were less training datasets, however it dropped as the number of training datasets increased. Together, these results indicate that CCIDeconv with the spatial features is able to deconvolute LR pairs when there is data sparsity. However, when training with a higher number of datasets, spatial information does not improve performance indicating CCIDeconv could be applied to datasets with no spatial information.

### 3.5. Deconvolution of non-spatial single cell data recovers previously studied communication patterns

We applied ScP procedure of CCIDeconv to a non-spatial scRNA-seq lung cancer dataset (Supplementary Methods 1.2.3) to assess if our model could recover known communication patterns. Having observed a median composite score of 0.71 during LOGO-CV training of the ScP procedure (Figure 4A-C, Supplementary Table 3), we tested the model on the single cell dataset. In the Lung Cancer and Brain (AD) datasets, region specific communication scores were poorly correlated with cell-level communication score ($R^2 \leq 0.7$), indicating that spatial information was altering the correlation (Supplementary Figure S9A,B). Therefore we used the remaining seven datasets as training datasets to predict communication patterns in the scRNA-seq lung cancer dataset. This ensured that CCIDeconv learned consistent and generalisable rules for deconvolution. We discovered high levels of cytoplasmic communication between the fibroblasts and mast cells (Figure 4D, Supplementary Figure S9C), which have been reported to occur through cell contact (Larsson-Callerfelt et al., 2021). Interestingly we discovered nucleus communication between the mononuclear phagocyte and malignant cells. The FN1-CD44 pair in these cell types had a higher nuclear score than cytoplasmic score. Although CD44 is expressed in the cytoplasm, it is well known to be translocated to the nucleus through endocytic pathways (Rios de la Rosa et al., 2019). As such we have demonstrated that CCIDeconv can be applied to single cell data, provided the training data is of sufficient quality.

## 4. Discussion

In this study, we identified differences in CCI depending on the subcellular location of the LRs. We found that the distribution of transcripts and subsequently their communication patterns vary within the subcellular regions across all tissue types. We developed a novel approach to deconvolve cell level communication scores to subcellular resolution. We have shown that CCIDeconv is accurate by assessing its performance by training nine sST datasets across multiple combinations of the other datasets. This framework was applied to non-spatial single cell data to deconvolve the communication scores to separate regions. Overall, we have shown that our method can effectively utilise sST data to deconvolute CCI to subcellular context for non-spatial scRNA-seq data.

In our analysis, we have used CellChatDB to characterize the LR pairs. CellChatDB is a popular database, as it has over 3000 curated LR pairs which include genes involved in contact based, extracellular matrix receptor interactions and secreted signalling pathways. sCCI involves ligands crossing the membrane of the cell and hence having such a diverse database is useful. However, CCIDeconv is not necessarily restricted by the database and can potentially use other databases such as OmniPath or CellPhoneDB to model sCCI.

Here we have focused on the task of deconvoluting CCI events stemming from either the nucleus or the cytoplasm of the cell. Of course, cells contain many different subcellular regions of interest not limited by the mitochondria, golgi body, or thin surface of the cell. While it is desirable to develop an analytical approach that can predict the origin of CCI events to all subcellular regions, many existing sST datasets are annotated to the resolution of the nucleus and cytoplasm only. With further refinement of subcellular annotations of such reference datasets, we would be equipped to extend CCIDeconv to regions beyond the nucleus and cytoplasm. One potential approach to ameliorate the lack of annotated subcellular reference data is to use the local gene expression abundances themselves as evidence for presence of cellular substructures, for example, subcellular regions enriched in mitochondrial genes may truly represent the mitochondria of the cell, and be annotated as such. This is an exciting avenue that could be explored in future work.

In the nine datasets we studied, we did not detect LR pairs unique to cross-region subcellular CCI. We speculate that this might be due to the limited gene coverage of the 10X Xenium panel, i.e., 5,000 genes rather than the whole transcriptome or the low expression of key signalling genes (antogonists or agonists). Therefore, expanding the panel to include additional genes involved in signal transduction would aid CCIDeconv to identify cross-region specific CCI.

Ideally one would interrogate at proteomics data to model CCI due to the proteins' role in cellular function. However subcellular spatial proteomics data is sparse and has limited protein coverage. Therefore, most methods, including CCIDeconv, use transcriptomics data as a proxy for assessing CCI. Hence the communication reflects co-expression rather than true protein binding and this may overestimate or miss functional interactions. We need to be cautious when interpreting the data due to lack of ground truth. Nevertheless, CCIDeconv can pinpoint where the CCI is occurring for a particular LR pair. Having this knowledge, one can utilise spatial proteomics or other experimental methods for the LR pair of interest to further confirm the results from CCIDeconv.

In conclusion, our findings highlight that CCI varies between subcellular regions. We developed a hierarchical model to attribute the communication scores into cytoplasm and nucleus. The results from our model could guide researchers to further refine and infer the subcellular location of CCI to understand disease mechanisms and signalling pathways.

### Conflicts of interest

The authors declare that they have no competing interests.

### Funding

The following sources of funding are gratefully acknowledged. S.G. was supported by an Australian Research Council DECRA Fellowship (DE220100964 to S.G.). P.P. was supported by the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (DAF2022–249319 and DAF2023-323340 to S.G.). J.Y.H.Y., was supported by Judith and David Coffey Research Fund, and National Health and Medical Research Council (NHMRC) Investigator Grant (APP2017023). R.J. was supported by the University of Sydney International Stipend and Tuition Fee Scholarship. The funding sources mentioned above had no role in the study design; in the collection, analysis and interpretation of data, in the writing of the manuscript and in the decision to submit the manuscript for publication.

### Data availability

No new data were generated in this study. All analyses were performed using publicly available datasets, as described in the Methods section, with links and references provided there.

### Code availability

Analysis and figure generation were performed using R (version 4.5.1) and Python (version 3.13.1). The model and scripts are available at https://github.com/SydneyBioX/CCIDeconv.

## Author contributions statement

S.G. and J.Y. conceived the study. R.J. developed the method and performed data analysis and wrote the initial draft of the manuscript under the supervision of J.Y., S.G., and P.P. All authors read and approved the final version of the manuscript.

## Acknowledgments

We thank our colleagues in the University of Sydney School of Mathematics and Statistics, Charles Perkins Centre, and Sydney Precision Data Science Centre for their support and intellectual engagement.

## REFERENCES

D. Aran, A. P. Looney, L. Liu, E. Wu, V. Fong, A. Hsu, S. Chak, R. P. Naikawadi, P. J. Wolters, A. R. Abate, A. J. Butte, and M. Bhattacharya. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol*, 20(2):163–172, Feb. 2019.

E. Armingol, A. Officer, O. Harismendy, and N. E. Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88, Nov. 2020.

S. Birk, I. Bonafonte-Pardàs, A. M. Feriz, A. Boxall, E. Agirre, F. Memi, A. Maguza, A. Yadav, E. Armingol, R. Fan, G. Castelo-Branco, F. J. Theis, O. A. Bayraktar, C. Talavera-López, and M. Lotfollahi. Quantitative characterization of cell niches in spatially resolved omics data. *Nat Genet*, 57(4):897–909, Apr. 2025.

A. Bock, B. T. Lobingier, M. Stoeber, and N. G. Tsvetanova. Chemical biology approaches to resolve the subcellular GPCR signaling landscape. *Nature chemical biology*, 21(8), Aug. 2025.

R. Browaeys, W. Saelens, and Y. Saeys. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods*, 17(2):159–162, Feb. 2020.

K. Chen, P. Iribarren, J. Hu, J. Chen, W. Gong, E. H. Cho, S. Lockett, N. M. Dunlop, and J. M. Wang. Activation of toll-like receptor 2 on microglia promotes cell uptake of alzheimer disease-associated amyloid beta peptide. *J Biol Chem*, 281(6):3651–3659, Feb. 2006.

J. C. Gardiner and E. Cukierman. Meaningful connections: Interrogating the role of physical fibroblast cell-cell communication in cancer. *Adv Cancer Res*, 154:141–168, Feb. 2022.

R. Gesztelyi, J. Zsuga, A. Kemeny-Beke, B. Varga, B. Juhasz, and A. Tosaki. The hill equation and the origin of quantitative pharmacology. *Archive for History of Exact Sciences*, 66(4):427–438, Mar. 2012.

V. Gramfort, B. Michel, O. Thirion, M. Grisel, and P. Blondel. *Machine Learning in Python Fabian Pedregosa, Gaël Varoquaux*, volume 12. Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay, 2011.

S. Jin, C. F. Guerrero-Juarez, L. Zhang, I. Chang, R. Ramos, C.-H. Kuan, P. Myung, M. V. Plikus, and Q. Nie. Inference and analysis of cell-cell communication using CellChat. *Nature Communications*, 12(1):1088, Feb. 2021.

S. Jin, M. V. Plikus, and Q. Nie. CellChat for systematic analysis of cell-cell communication from single-cell transcriptomics. *Nat Protoc*, 20(1):180–219, Jan. 2025.

A. Larsson-Callerfelt, M. Bagher, O. Rosmark, L. Elowsson-Rendin, A. Nybom, S. Wasserstrom, C. Müller, X. Zhou, G. Dellgren, O. Hallgren, L. Bjermer, and G. Westergren-Thorsson. Crosstalk between lung fibroblasts and mast cells is modified by alveolar extracellular matrix and influences epithelial migration. In *03.03 - Mechanisms of lung injury and repair*, page 102. European Respiratory Society, Mar. 2021.

R. Li and X. Yang. De novo reconstruction of cell interaction landscapes from single-cell spatial transcriptome data with DeepLinc. *Genome Biol*, 23(1):124, June 2022.

A. Lorico, M. F. Santos, J. Karbanová, and D. Corbeil. Extracellular membrane particles en route to the nucleus - exploring the VOR complex. *Biochem. Soc. Trans.*, 53(3):529–546, June 2025.

D. McCarthy, K. Campbell, A. Lun, and Q. Wills. lt;igt;scater:lt;/igt;pre-processing, quality control, normalisation and visualisation of single-cell rna-seq data in r. *bioRxiv*, 2016. doi: 10.1101/069633. URL `https://europepmc.org/article/PPR/PPR7011`.

J. A. Y. Y. S. D. T. N. A. A. G. M. S. D. P. A. T. Monica Nagendran, Jerald Sapida. 1457 visium HD enables spatially resolved, single-cell scale resolution mapping of FFPE human breast cancer tissue. *Journal for ImmunoTherapy of Cancer*, 11(Suppl 1), Nov. 2023.

F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Comput. Stat.*, Mar. 2022.

B. Z. Peters Couto, N. Robertson, E. Patrick, and S. Ghazanfar. MoleculeExperiment enables consistent infrastructure for molecule-resolved spatial omics data in bioconductor. *Bioinformatics*, 39(9), Sept. 2023.

J. M. Rios de la Rosa, P. Pingrajai, M. Pelliccia, A. Spadea, E. Lallana, A. Gennari, I. J. Stratford, W. Rocchia, A. Tirella, and N. Tirelli. Binding and internalization in receptor-targeted carriers: The complex role of CD44 in the uptake of hyaluronic acid-based nanoparticles (siRNA delivery). *Adv. Healthc. Mater.*, 8(24):e1901182, Dec. 2019.

J. Su, Y. Song, Z. Zhu, X. Huang, J. Fan, J. Qiao, and F. Mao. Cell–cell communication: new insights and clinical implications. *Signal Transduction and Targeted Therapy*, 9(1):196, Aug. 2024.

J. Tanevski, R. O. R. Flores, A. Gabor, D. Schapiro, and J. Saez-Rodriguez. Explainable multiview framework for dissecting spatial relationships from highly multiplexed data. *Genome Biol*, 23(1):97, Apr. 2022.

The Human Protein Atlas. The human protein atlas. `https://www.proteinatlas.org`. Accessed: 2026-2-10.

K. Troulé, R. Petryszak, B. Cakir, J. Cranley, A. Harasty, M. Prete, Z. K. Tuong, S. A. Teichmann, L. Garcia-Alonso, and R. Vento-Tormo. CellPhoneDB v5: inferring cell–cell communication from single-cell multiomics data. *Nature Protocols*, 20(12):3412–3440, Mar. 2025.

D. Türei, T. Korcsmáros, and J. Saez-Rodriguez. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods*, 13(12):966–967, Nov. 2016.

M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg,

L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Pontén. Proteomics. tissue-based map of the human proteome. *Science*, 347(6220):1260419, Jan. 2015.

M. V. Vargas, L. E. Dunlap, C. Dong, S. J. Carter, R. J. Tombari, S. A. Jami, L. P. Cameron, S. D. Patel, J. J. Hennessey, H. N. Saeger, J. D. McCorvy, J. A. Gray, L. Tian, and D. E. Olson. Psychedelics promote neuroplasticity through the activation of intracellular 5-HT2A receptors. *Science*, 379(6633):700–706, Feb. 2023.

Y. Zhao, Y. Li, Y. He, J. Wu, Y. Liu, X. Li, Z. Li, Q. Yuan, J. Li, X. Zhang, X. Hu, G. Wu, Y. Qiu, J. Yuan, X. Huang, M. Xu, M. Yang, J. Guo, T. Cao, Q. Chen, D. Li, J. Zhang, X. Zhou, X. Jiang, F. Zhu, X. Dong, R. Xiang, H. Pan, L. Han, Z. Deng, H. Deng, Y. Zhang, M. Liu, Q. Wu, G. Wang, J. Zhai, W. Tan, X. Liu, Z. Wang, S. Li, T. Duan, L. Liu, A. Chen, H. Liu, C. Chen, S. Liao, and X. Xu. Stereo-seq V2: Spatial mapping of total RNA on FFPE sections with high resolution. *Cell*, 188(23):6554–6571.e21, Nov. 2025.

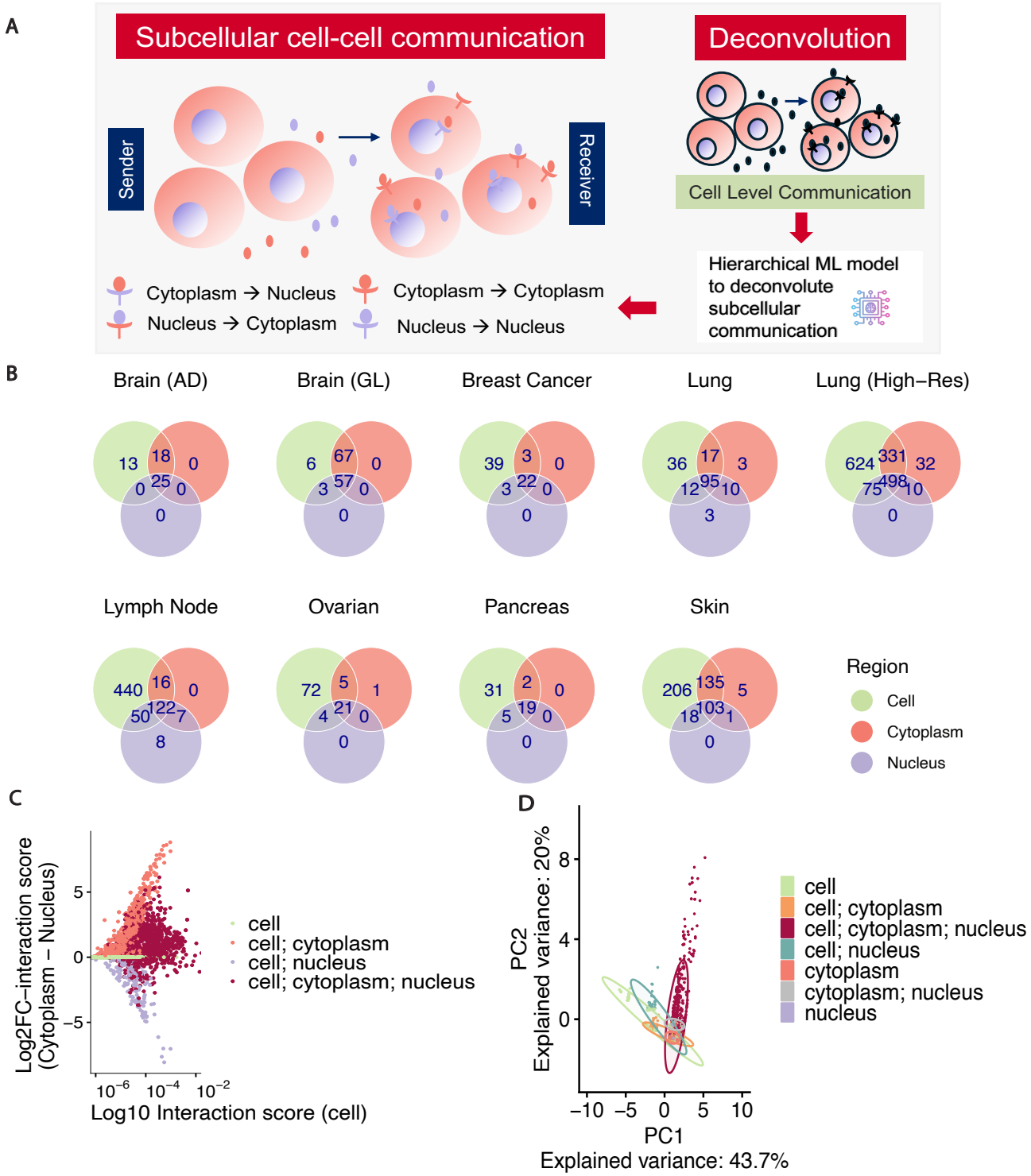


**Figure 1** Differences in subcellular communication across nine different datasets. A. Overview of subcellular and cell-cell communication, and quantification of communication scores across the subcellular regions. B. Significant interactions detected using CellChat are different between the cytoplasm and nucleus across the nine datasets. C. Log2 Fold Change differences in the LR communication score detected between the cell and its subcellular regions. D. PCA of CellChat-derived communication score components.

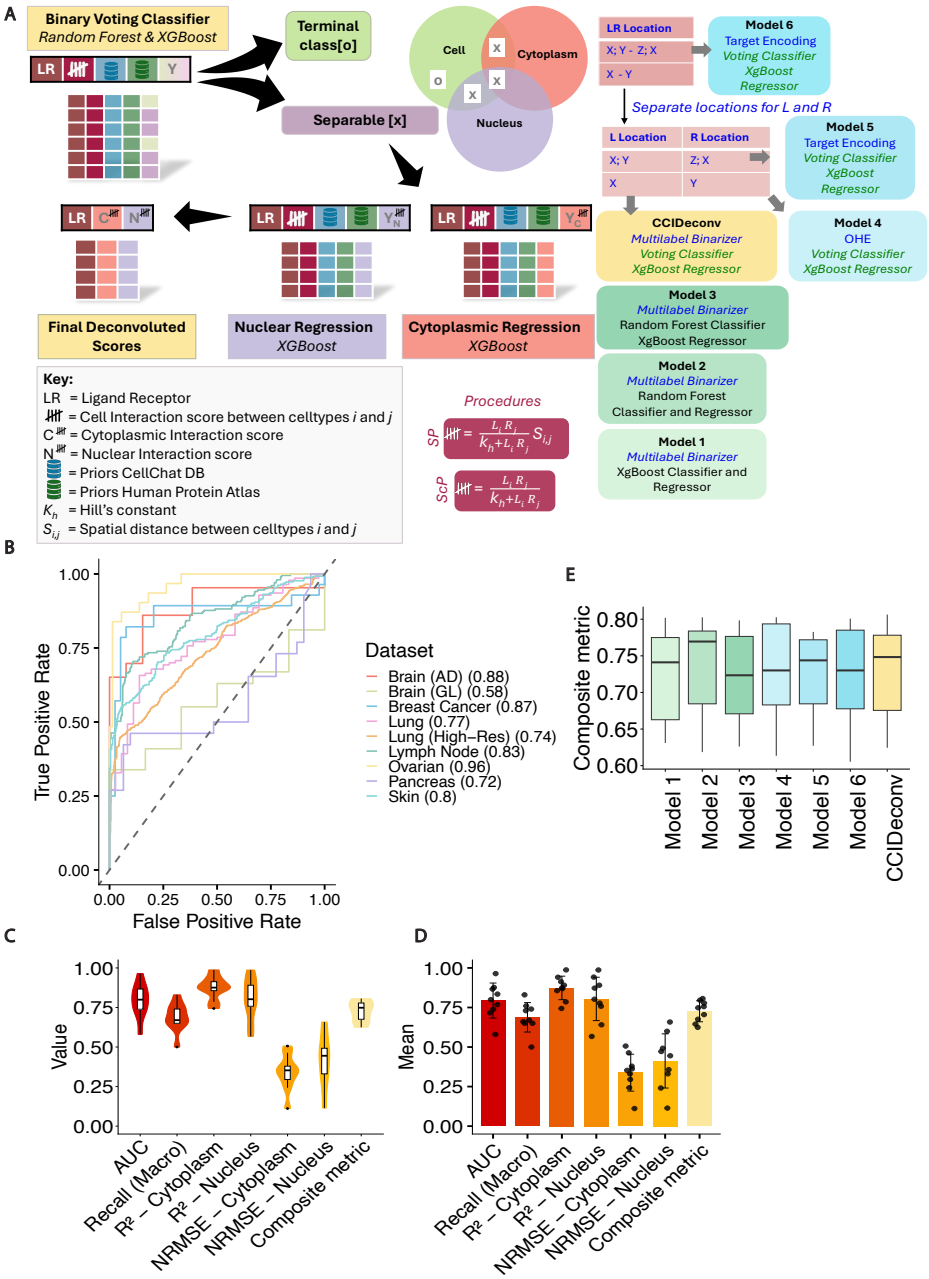


**Figure 2** Training and evaluation of CCIDeconv, a hierarchical model for predicting subcellular cell communication scores. A. Schematic of the hierarchical model architecture of CCIDeconv, combining a classifier and a regressor to predict communication scores in subcellular regions. B. ROC curves showing the classification performance of the model on nine datasets, with the area under the curve (AUC) for each dataset indicated in the legend. C. Violin plots showing the distribution of performance metrics. D. Barplots representing the mean of performance metrics across datasets with standard error bars. E. Composite metrics comparing different base learners and feature engineering approaches.

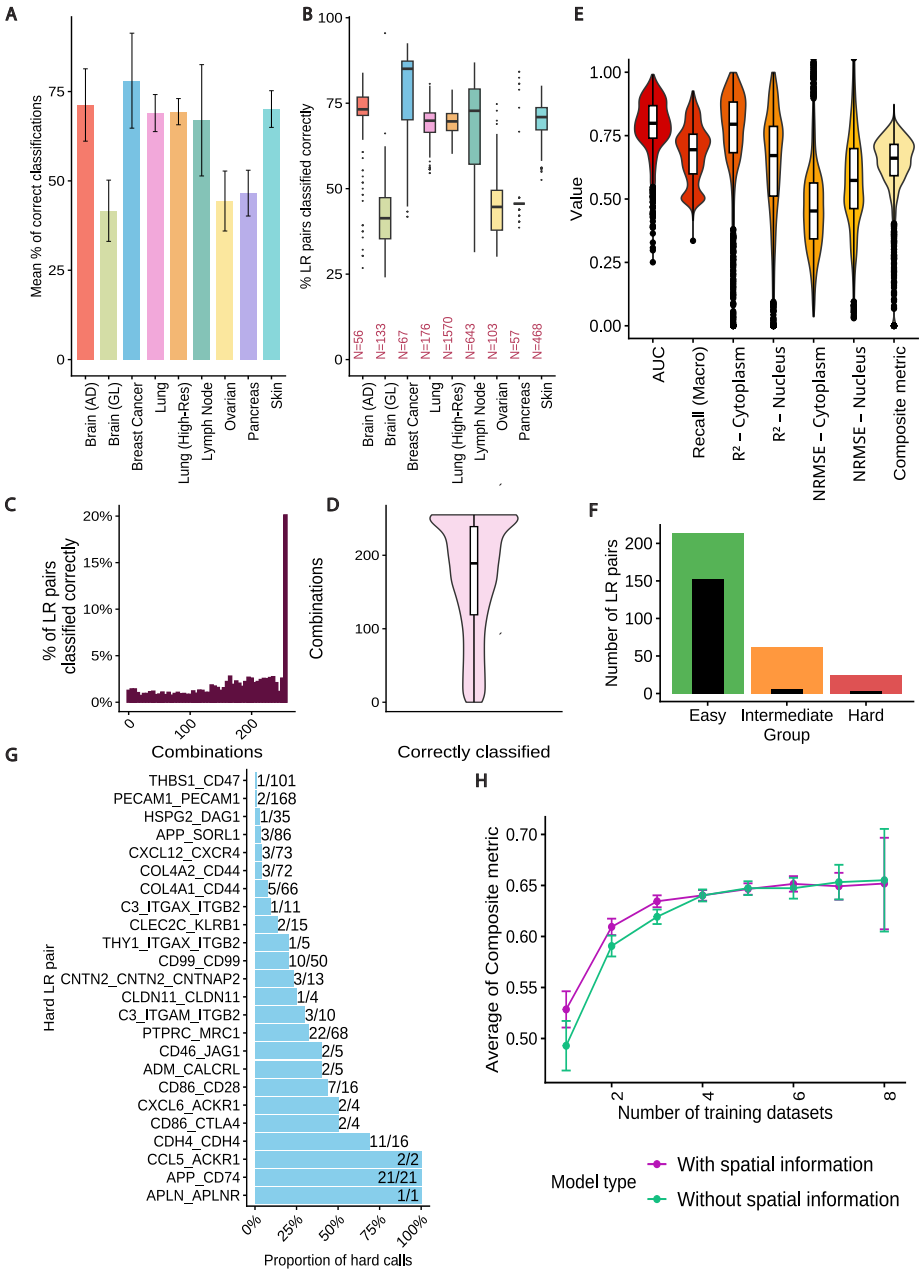


**Figure 3** Assessing model stability by training each dataset with varying combinations of the remaining datasets. A, B. Mean percentage (A) and distribution (B) of significant ligand–receptor (LR) pairs correctly classified when each dataset was trained with 255 different combinations of the other eight datasets. C, D. Number of training combinations in which each significant LR pair across the nine datasets was correctly classified: (C) histogram of correctly classified LR pairs; (D) violin plots showing the distribution of the number of combinations in which each LR pair was correctly classified. E. Violin plots of performance metrics across the 255 training combinations for all nine datasets (outliers with cytoplasmic and nuclear $R^2 \leq 0$ are not shown). F. Counts of LR pairs consistently assigned to difficulty groups. Bars represent the number of significant LR pairs that were classified as Easy ($\geq$ 75%), Intermediate (25–75%), or Hard ($\leq$ 25%) based on the proportion of training combinations in which they were correctly classified across all datasets. Gray overlays indicate LR pairs that were always assigned to the same difficulty group across all combinations. G. Bar plot of LR pairs classified as Hard; fractions shown to the right of each bar indicate the proportion of combinations in which each LR pair was labeled as Hard. H. Composite metrics for models trained with SP and ScP procedure across different dataset combinations. A total of 4,590 models were trained (each of the nine datasets trained across 255 combinations of other datasets with both the SP and ScP procedures); plots show mean composite metrics with error bars.

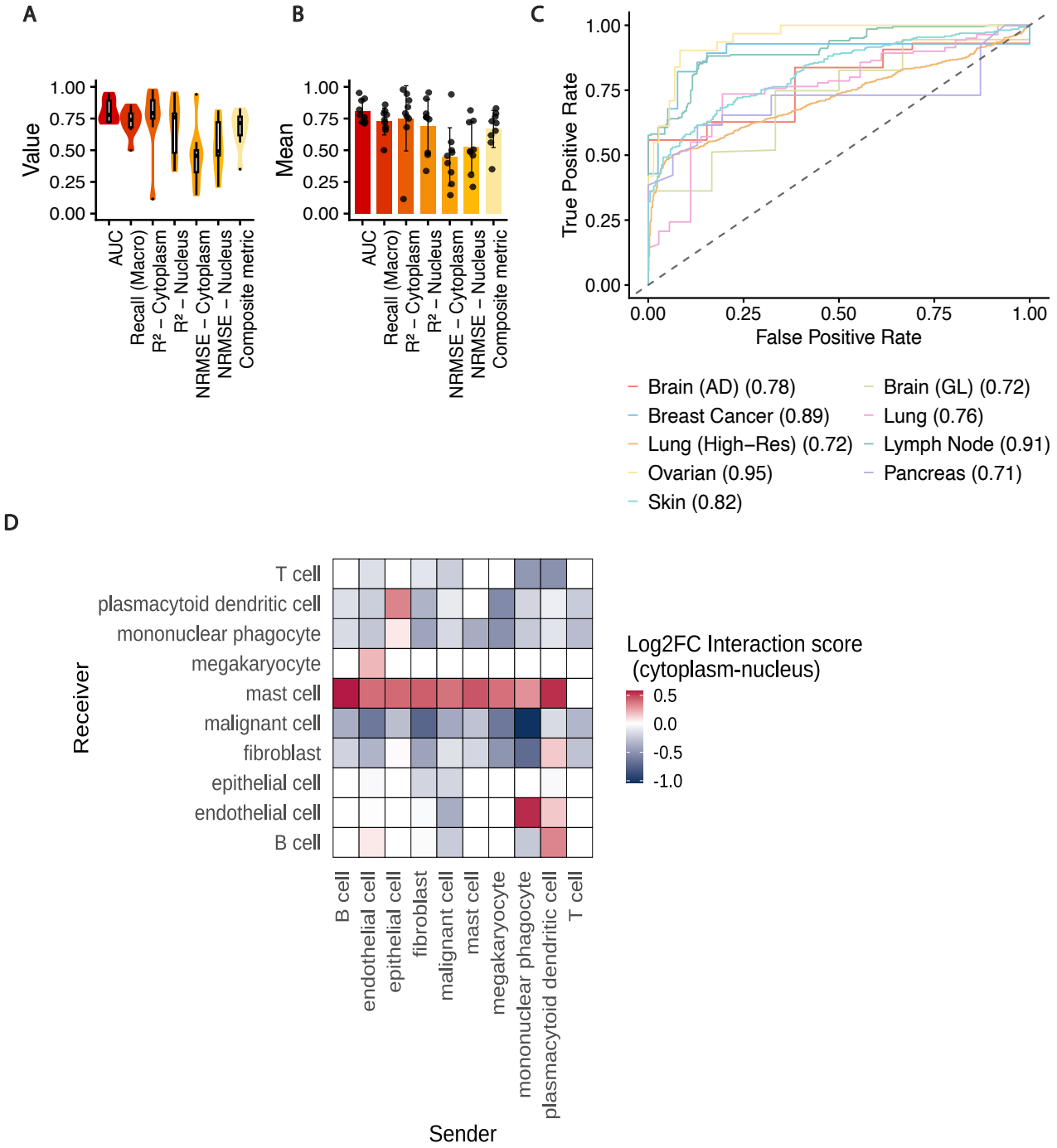


**Figure 4** Training the CCIDeconv through ScP procedure to test applicability to single cell data. A - C Performance metrics on training the model without the spatial features: (A) Violin plots representing the distribution of the metrics across the datasets (B) Bar plots representing the mean across datasets with standard error bars and (C) ROC curve. D. Heatmap showing the difference in the communication scores when the model is applied on a single cell data.