UC Irvine
UC Irvine Electronic Theses and Dissertations

Title
Computational Modeling of Cell-cell Communication Patterns from Single-Cell and Spatial Transcriptomic Data

Permalink
https://escholarship.org/uc/item/8cg8f5mv

Author
Wang, Xinyi

Publication Date
2025

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA,
IRVINE

Computational Modeling of Cell-cell Communication Patterns from Single-Cell and Spatial Transcriptomic Data

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematics

by

Xinyi Wang

Dissertation Committee:
Professor Qing Nie, Chair
Professor Jun F. Allard
Professor Christopher E. Miles

2025

Chapter 1 © 2023 Seminars on Cancer Biology
Chapter 2 © 2022 Frontiers in Genetics
Chapter 3 © 2024 Nature Communications
Chapter 4 © 2024 Nature Communications
All other materials © 2025 Xinyi Wang

TABLE OF CONTENTS

Page

LIST OF FIGURES **iv**

ACKNOWLEDGMENTS **vi**

VITA **vii**

ABSTRACT OF THE DISSERTATION **ix**

1 The Promising Application of Cell-cell Communication Analysis from Single-cell and Spatial Transcriptomics **1**
1.1 Introduction . . . 1
1.2 Results . . . 5
1.2.1 CCC methods . . . 5
1.2.2 Downstream analysis of CCC . . . 15
1.2.3 Experimental Validation . . . 17
1.2.4 Analysis of CCC in cancer . . . 18
1.2.5 Future Directions . . . 20
1.3 Discussion . . . 22

2 Analyzing Diversity of Cell-cell Communication Networks via Single-cell Transcriptomics **24**
2.1 Introduction . . . 24
2.2 Results . . . 26
2.2.1 Analysis of CCC in different COVID-19 Conditions . . . 26
2.2.2 Analysis of CCC in MIF pathway . . . 30
2.2.3 Cell type diversity of CCC in MIF pathway . . . 33
2.2.4 A random walk inspired node-to-node information flow in CCC graphs in MIF pathway . . . 35
2.2.5 Canberra node-to-node similarity in CCC graphs in MIF pathway . . . 37
2.2.6 Weighted cosine node-to-node similarity in CCC graphs in MIF pathway 39
2.2.7 Diversity of CCC networks in the MIF pathway . . . 40
2.3 Discussion . . . 40
2.4 Method Overview . . . 43
2.4.1 Data Preparation . . . 43
2.4.2 CCC Analysis by CellChat . . . 44

2.4.3 Measures of Graph Features 45

3 Detect Hierarchical Structures in Cell-cell Communications via Single-cell and Spatial Transcriptomics Data 50
3.1 Introduction 50
3.2 Methods 53
3.2.1 Overview of CrossChat 53
3.2.2 CrossChatH: hierarchical clustering 57
3.2.3 CrossChatT: CCC Tree search 65
3.2.4 CrossChatS: Incorporating spatial information 66
3.2.5 Incorporation of CellChat and COMMOT 66
3.2.6 Validation of CrossChat using simulated dataset and COVID-19 dataset 67
3.2.7 Robustness of CrossChatH 71
3.3 Discussion 72

4 Applications of CrossChatH and CrossChatT to Single-cell and Spatial Transcriptomics Data 74
4.1 Introduction 74
4.2 Results 75
4.2.1 CrossChatH identifies hierarchical clusters and specific interactions in COVID-19 patients 75
4.2.2 CrossChatT identifies local hierarchical signaling structures driven by ligand-receptor interactions in mouse embryonic skin 80
4.2.3 CrossChatS reveals hierarchical clusters and ligand-receptor interactions in spatial datasets 84
4.2.4 Comparison between CrossChat and other CCC detection methods . 90
4.3 Discussion 93

Bibliography 95

Appendix A Additional file for Chapter 2 104

Appendix B Additional File for Chapter 3 113

Appendix C Additional File for Chapter 4 117

LIST OF FIGURES

Page

1.1 The biological components of cell-cell interactions that can be inferred from single-cell and spatial data 4

2.1 CCC analysis using CellChat. 29
2.2 CCC analysis using CellChat focused on the MIF pathway 32
2.3 Distribution of cell type diversity statistics in different COVID-19 conditions are visualized 36
2.4 Distribution of cell-type-to-cell-type features 38
2.5 Network variability based on four distance measures 41

3.1 Hierarchical structures in cell-cell communications 56
3.2 Method overview of CrossChatH and CrossChatT. 64
3.3 Validation of CrossChatH 70

4.1 COVID-19 PBMC analysis using CrossChatH. 79
4.2 Applications of CrossChatT to mouse skin cells during wound healing. . . . 83
4.3 Applications of CrossChatS to spatial datasets. 89

A.1 CellChat analysis of all pathways after downsampling 105
A.2 CellChat analysis of MIF pathway after downsampling 106
A.3 Visualization of information flow using nonlinear embeddings 107
A.4 Distribution of ten cell types diversity statistics in different COVID-19 conditions are visualized 108
A.5 Distribution of node-to-node information flow statistic of ten cell types in MIF pathway in different COVID-19 conditions and comparison of nonlinear projection of information flow proximity matrices of three cell types 109
A.6 Distribution of node-to-node Canberra similarity of ten cell types in different COVID-19 conditions are visualized 110
A.7 Distribution of node-to-node weighted cosine similarity of ten cell types in different COVID-19 conditions are visualized 111
A.8 Network variability based on four distance measures in ten cell type networks 112

B.1 Robustness test of CrossChatH with respect to different parameters 114

C.1 Application of CrossChatH to PBMC cells sampled from COVID-19 patients 117
C.2 Demonstration of spatial incorporation for CrossChatH-S 118

C.3 Application of CrossChatH-S to Stereo-seq mouse embryo at day E16.5 data 119
C.4 Cell-cell communications on Stereo-seq mouse embryo at day E16.5 and multiscale similarity of ligand-receptor interactions 120
C.5 Cell-cell communication inferred from 10x Visium mouse skin. 122
C.6 Comparison of CrossChat with existing CCC methods 124
C.7 Comparison of CrossChat-S with existing spatial CCC methods 125
C.8 Comparison of CrossChatH analysis with different CCC packages 127
C.9 Comparison of CrossChatT analysis with different CCC methods 129
C.10 Hierarchical clustering on CCC calculated using single-cell level inference tools 131

ACKNOWLEDGMENTS

I would like to express my most sincere gratitude to my advisor and my committee chair, Professor Qing Nie, who provides insightful suggestions and discussions, and provides the freedom to explore interesting directions. I sincerely thank his encouragement, understanding and support throughout the years.

I would also like to thank my committee members, Professor Jun Allard and Professor Chris Miles, for their support, encouragement and insightful comments.

I would like to thank all members of Nie Lab. In addition, I thank Axel Almet for discussions and suggestions in improving the manuscripts.

Chapter 1 of this thesis is an adaptation of the material as it appears in [105], used with permission from Seminars in Cancer Biology. The coauthors listed in this publication are Axel Almet and Qing Nie. Qing Nie directed and supervised research which forms the basis for the thesis.

Chapter 2 of this thesis is an adaptation of the material as it appears in [104], used with permission from Frontiers in Genetics. The coauthors listed in this publication are Axel Almet and Qing Nie. Qing Nie directed and supervised research which forms the basis for the thesis.

Chapter 3 and 4 of this thesis is an adaptation of the material as it appears in [100], used with permission from Nature Communications. The coauthors listed in this publication are Axel Almet and Qing Nie. Qing Nie directed and supervised research which forms the basis for the thesis.

I would also like to thank the funding sources that have supported me during my time at the University of California, Irvine. These include the UCI Mathematics department, National Science Foundation and the Simons Foundation, as well as on-campus centers that provided the facilities and resources to perform the research presented, including the NSF-Simons Center for Multiscale Cell Fate Research, the Center for Complex Biological Systems.

In addition, the work in this dissertation was funded by National Science Foundation grant DMS176372 and CBET2134916, the National Institutes of Health grants R01AR079150, R01DE030565, R01GM152494, and The Simons Foundation Grant (594598 Q.N.).

Finally, I thank my partner Yucheng for his enduring love, trust, and support. And to my parents, who set me off on the road years ago and provide continuous care and support.

VITA

Xinyi Wang

EDUCATION

Doctor of Philosophy in Mathematics **2025**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Mathematics **2019**
University of California, San Diego *La Jolla, CA*

RESEARCH EXPERIENCE

Graduate Research Assistant **2021–2025**
University of California, Irvine *Irvine, CA*

TEACHING EXPERIENCE

Teaching Assistant **2020–2025**
University of California, Irvine *Irvine, CA*

PUBLICATIONS

1. Xinyi Wang, Axel Almet, Qing Nie, 2021, Analyzing network diversity of cell–cell interactions in COVID-19 using single-cell transcriptomics. *Frontiers in Genetics 13* (2022): 948508.

2. Xinyi Wang, Axel Almet, Qing Nie, 2022, The promising application of cell-cell interaction analysis in cancer from single-cell and spatial transcriptomics. *Seminars in Cancer Biology* 13 (2023): 948508.

3. Xinyi Wang, Axel Almet, Qing Nie, 2023, Detecting global and local hierarchical structures in cell-cell communication using CrossChat. *Nature Communications* 15.1 (2024): 1-16.

SOFTWARE

CrossChat `https://github.com/Xinyiw28/CrossChat`
Python algorithms that detects hierarchical structures in cell-cell communications from scRNA-seq and Spatial Transcriptomic (ST) datasets

ABSTRACT OF THE DISSERTATION

Computational Modeling of Cell-cell Communication Patterns from Single-Cell and Spatial Transcriptomic Data

By

Xinyi Wang

Doctor of Philosophy in Mathematics

University of California, Irvine, 2025

Professor Qing Nie, Chair

Cells work in concert to perform functions in a multicellular organism. This coordination occurs through cell–cell communication (CCC), where cells exchange molecular messages. CCC is fundamental to numerous biological processes, such as cell growth, division, differentiation, tissue and organ development, and disease progression.

Recent technological breakthroughs in single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics have revolutionized the study of CCC. While a rich number of methods have been developed to infer CCC, the underlying patterns and structures within cell communication events remain insufficiently understood. In Chapter 1, we reviewed the vast number of methods that have been developed to infer CCC based on scRNA-seq or spatial transcriptomics datasets. In Chapter 2, we analyzed the diversity of CCC networks within different disease conditions. We proposed measures at 3 scales, including node-level, edge-level, and graph-level, that comprehensively analyzes CCC networks at different levels. In Chapter 3, we proposed CrossChat, which composed of two core methods, CrossChatH and CrossChatT, to detect and analyze hierarchical structures within CCC with respect to ligand and receptor expression in two complementary angles. In Chapter 4, we applied CrossChat to scRNA-seq and spatial transcriptomics datasets to illustrate its functionalities and showcase

its novelty over existing CCC methods.

Chapter 1

The Promising Application of Cell-cell Communication Analysis from Single-cell and Spatial Transcriptomics

1.1 Introduction

Cells coordinate to perform functions as a multicellular organism. Such coordination can be achieved through cell-cell interactions (CCC)—also known as cell-cell communication or cell signaling—where cells communicate with nearby cells by sending and receiving molecular messages. Within multicellular organisms, there are distinct groups of specialized cells with distinct cell functions, which facilitate different types of CCC. These range of interaction behaviors are crucial to many biological processes, including cell growth, division, differentiation, tissue or organ development, and disease progression [4, 65, 92]. For example, in the

context of cell differentiation, T follicular helper cells, a subset of CD4+ T cells, are found to secrete IL-4 and IL-21 ligands that promote B cell proliferation and B cell differentiation into plasma B cells or germinal center B cells [30]. In embryonic development, WNT signaling are found to play diverse roles in cell fate determination, differentiation, proliferation and apoptosis [94]. In disease progression, some ligands and receptors are found to inhibit interactions involved in facilitating the immune responses. For instance, cytotoxic T lymphocyte-associated antigen-4 (CTLA-4) was found to downregulate immune responses and is closely related to tumor progression (Ma et al. 2021). In particular, CCC is important for cancer studies. Many cellular processes that are crucial for the development of cancer are regulated by CCC, including cell growth and division, cell apoptosis, cell motility and invasion, angiogenesis, inflammation and immune suppression [45, 85, 46].

There are two components of CCC. One is intercellular signaling, which occurs between cells and at the cell membrane (Figure 1.1a). The other is intracellular signaling, which is the downstream response to intercellular signaling and takes place inside the cells (Figure 1.1b). Intercellular signaling consists of sender cells secreting signaling molecules, called ligands, into the extracellular space. Each ligand can then only bind to a certain set of proteins located on the membrane of possible receiver cells called receptors. When a sufficient number of ligands have reached and binded to their appropriate receptors, different transcription factors and consequently target genes within the cell are activated downstream. There are four types of cell communication: autocrine, paracrine, endocrine and juxtacrine. Autocrine signaling is defined as a cell receiving a signal secreted by itself. For example, it was found that autocrine TGFb signaling maintains self-renewal in human embryonic stem cells [74]. Paracrine signaling is defined as a cell receiving a signal secreted by other nearby cells. For example, paracrine Interleukin-1 signaling from from carcinoma cells induces cytokine secretion in mesenchymal stem cells [81]. Juxtacrine signaling requires direct physical contact between cells, where signals are secreted across gap junctions between cells in contact. An example of this type of signaling is the interaction between a membrane-bound ligand Delta

and the cell-surface receptor Notch [19]. Signals secreted in endocrine signaling travel a long distance from sender cells through the circulatory system to reach the potential receiver cells. The release of hormones from glands and travel through the bloodstream to reach distant body sites is an example of endocrine. However, due to the far spatial distance, scRNA-seq and ST data do not provide enough information to infer endocrine signaling.

Single-cell RNA sequencing (scRNA-seq) and spatial transcriptomic (ST) technologies are rapidly developing, which enable the profiling of biological tissue at unprecedented genomic depth. The gene expression profile of tens of thousands of genes, many of which are related to CCC, can be profiled for thousands and potentially millions of cells [90]. Being able to capture such rich genomic information offers a great opportunity to change how CCC can be investigated in different tissues of different species. For known ligand-receptor interaction pairs, based on the assumption that higher levels of relevant ligand and receptor gene expression reflect a higher possibility of CCC occurring, one can use the expression of ligand and receptor genes in possible sender and receiver cells to infer CCC. A great number of bioinformatics tools have been developed recently to model and analyze CCC between and within cells based on gene expression data obtained from non-spatial single-cell and spatial transcriptomics data—and many more continue to be developed. Using these computational tools, we can infer an approximate landscape of CCC from scRNA-seq data (or ST data), which can advance our understanding of CCC mechanisms in different biological systems. In particular, by analyzing cancer datasets using CCC tools, we can gain a better understanding of the interacting cell groups and the ligand and receptor pairs involved in the progression of cancer. Indeed, there is a rapidly growing number of applications where these tools have been used to reveal important and novel CCCs from scRNA-seq studies of cancer [58, 54, 75, 43]. Most CCC tools consist of two components: a ligand-receptor database which contains the possible ligand-receptor pairs and a computational model to calculate the likelihood of CCC based on the expression values of the ligand and receptor genes. Some CCC tools, such as CellChat [54], include functionality to visualize the CCC networks directly. In this paper,

we will briefly review the currently available computational tools that model CCC based on gene expression data obtained from non-spatial scRNA-seq or ST data. We discuss current ways to perform downstream analysis after inferring CCC, how CCCs are validated, the applications of CCC results in cancer, and possible future directions.

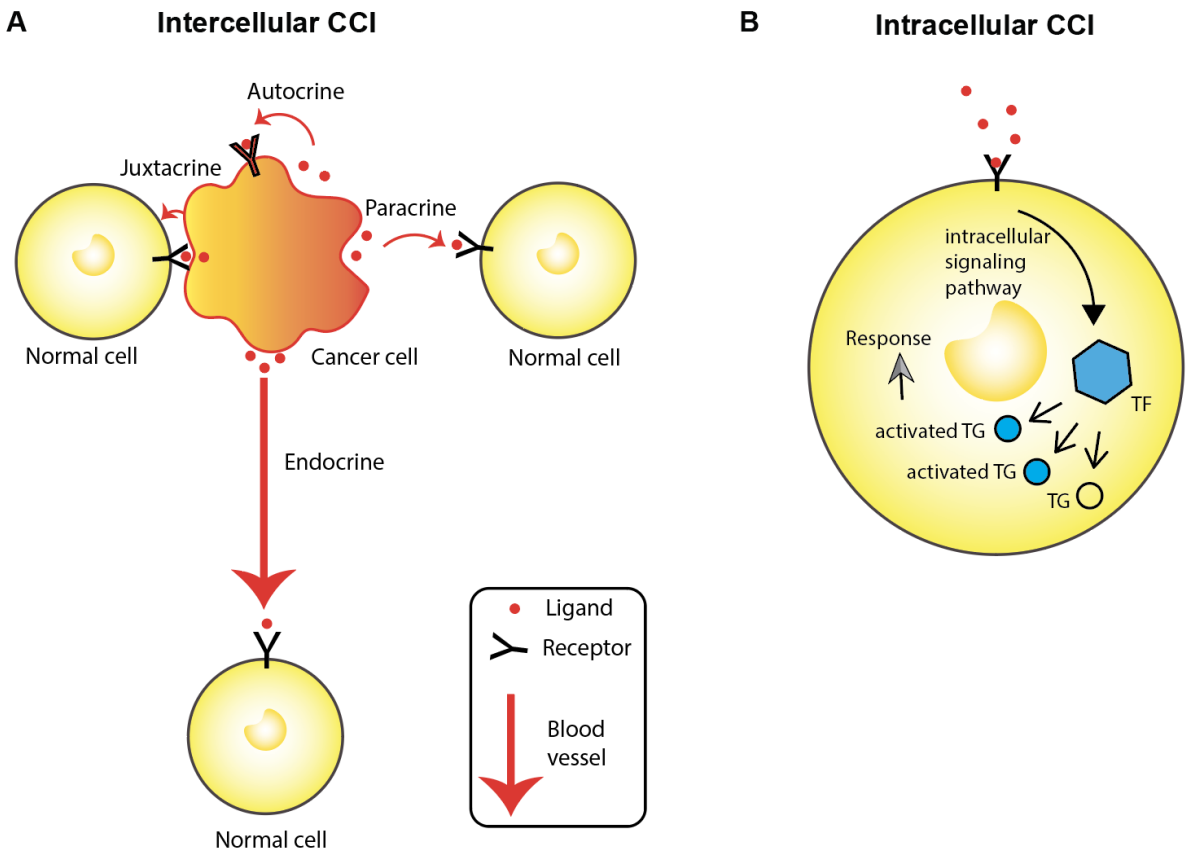


Figure 1.1: **The biological components of cell-cell interactions that can be inferred from single-cell and spatial data.** (a) The different types of intercellular CCC, which include autocrine, paracrine, and juxtacrine signaling. Autocrine signaling occurs when a cell receives the same signals secreted by itself. Paracrine signaling occurs when a cell receives a signal secreted by nearby cells. Juxtacrine signaling occurs a cell receives a signal from a directly adjacent cell through physical contact. (b) Intracellular CCC is the downstream response to intercellular signaling that takes place inside cells. When a sufficient number of ligands bind to associated receptors, the receptors are activated. The signal received by receptors will be converted into an intracellular signal and activate transcription factors and target genes.

1.2 Results

1.2.1 CCC methods

Many computational CCC tools have been developed based on either non-spatial scRNA-seq data or have used ST data to constrain potential interactions. The CCC tools usually take preprocessed data, which includes normalization by library depth and log-transformation, as input. These tools cover both the calculation of intercellular interactions and intracellular interactions. While these tools have the common goal of modeling CCC, they are based on different computational strategies and biological assumptions. We will describe the currently available tools in detail below.

Intercellular CCC methods

In this section, we briefly describe the categories of computational methods that CCC tools use to model the different components of cell signaling.

Intercellular CCC methods are used to infer potential interactions between cells. More specifically, they try to capture the interaction strength, which is assumed to reflect the probability of an interaction occurring, between associated ligands and receptors.

Threshold based methods: One type of CCC method removes insignificant interactions based on individual ligand and receptor expression levels, where ligands and receptors are only considered if their expression levels exceed a predetermined threshold in their respective sender and receiver cell types. After thresholding, only ligand-receptor pairs where both the ligand and receptor are retained for CCC inference. These types of methods output a binary CCC score. One example of this type of method is CellTalker [25], which only considers ligands and receptors with non-zero expression in more than 5% of cells for inference, or

ligands and receptors with non-zero expression in more than 3% of certain cell type groups for specific interactions.

Differential-expression-based methods Differential-expression-based: CCC methods infer significant interactions by first identifying differentially expressed ligands and receptors using statistical models. The differentially expressed ligands and receptors are then compared to existing ligand-receptor pairs in a curated CCC database. The output of these CCC methods are generally interpreted as a binary score for all ligand-receptor pairs. PyMINEr is an example of one such CCC method [95]. It first finds the significantly enriched ligands and receptors in each cell type based on both analysis of variance (ANOVA) and z-score enrichment, and then cross-references ligand-receptor pairs with the StringDB interaction list to only include direct binding pairs. Another example of one such CCC method is iTALK, which allows for the use of one of several differential expression methods for single-cell applications to find differentially expressed ligands and receptors.

Permutation-based methods: Permutation-based CCC methods measure the specificity of an interaction between two cell types. There are two types of permutation tests considered: tests that permute either gene labels or cell type labels. Gene label permutation tests measure how high the observed ligand (receptor) expression levels are compared with a "null distribution" obtained from randomly selected genes. Cell type label permutation tests measure how high the ligand and receptor expression level in the considered cell types are compared to randomly assigned cell type labels.

One example of a method that performs gene label permutation tests is scSeqComm [8]. Given a fixed cell type, it first randomly resamples genes from all genes and computes the distribution of the average gene expression of a randomly resampled gene. The distribution of average gene expressions observed by chance is then approximated using a Gaussian distribution. The score of the ligand (receptor) in the fixed cell type is computed as the probability of observing lower values from the approximated distribution than the average

gene expression of the ligand (receptor) of interest. The ligand-receptor score is calculated as the minimum of the ligand and receptor score.

CellPhoneDB [35] is an example of a method that performs cell type label permutation. Assuming cell type annotations have been provided, CellPhoneDB calculates an enrichment score of ligand-receptor interactions between two cell types based on the gene expression of ligand and receptor in sender and receiver cell types, respectively. This enrichment score is calculated as the minimum of the average gene expression of ligand and receptor in their respective cell types. Next, by randomly permuting the cell type labels of each cell a large number of times, a null distribution of enrichment score of ligand-receptor interaction is calculated. Then the interaction score of a ligand-receptor pair is calculated as the proportion of enrichment scores that are higher than the actual computed enrichment score.

Coexpression-based methods: The majority of CCC methods analyzed in this paper are based on the measuring the simultaneous coexpression of ligands and receptors. These types of methods can be further divided into three subcategories based on their calculation formula; that is, whether they are sum based, product based, or correlation based.

(i). Sum-based coexpression: Sum-based coexpression methods infer CCC based on a calculated sum of certain features of ligand and receptor expression. For example, for a given ligand-receptor interaction, CellCall infers intercellular communication by calculating the Euclidean norm of a vector consisting of the normalized ligand and receptor expression values, which is then weighted by an activity score of associated transcription factors [112]. CytoTalk is another method that uses a sum-based method to calculate the ligand-receptor interaction strength [50]; possible interactions are weighted by multiplying the sum of features of ligand and receptor expression with a "non-self-talk" score, which is calculated using the mutual information of ligand and receptor expression in the sender group and receiver groups, respectively.

(ii). Product-based coexpression: Product-based coexpression methods, as their name suggests, infer CCC based on the product of ligand and receptor gene expression. CellChat [54] is one such popular method. CellChat considers the expression levels of ligands and receptors in respective sender and receiver cell groups, respectively. When considering multi-units of ligands or receptors, CellChat uses the geometric mean of each gene to approximate the average expression level. The communication probability between two cell groups for a given ligand-receptor pair is then defined as a product based on mass action kinetics, where the "base" ligand-receptor score, calculated as a normalized product of ligand and receptor expression, which is then weighted by the average expression of known agonists and antagonists.

Product-based coexpression methods are popular for CCC inference. Other product-based methods include: NATMI [49], SingleCellSignalR [16], ICELLNET [72], scConnect [52], CSOmap [78], SoptSC [102], and Connectome [76]. Each method uses its own "product-based" formula to infer ligand-receptor interaction strength.

(iii). Correlation-based coexpression: Correlation-based coexpression methods infer CCC based on statistical correlation between gene expression in cell groups of interest. These methods assume that correlation in expression of two genes correspond to regulation by a common signaling mechanism. REMI calculates the Pearson correlation between ligand and receptor gene expressions [109] to represent the interaction.

Intracellular CCC methods

Intracellular CCC methods model the interaction process within cells. Some intracellular CCC methods specifically model downstream networks containing interactions from receptors to transcription factors, and from transcription factors to downstream target genes, while others model interactions among all possible intracellular genes within the cells.

Fisher's Exact test: To construct the intracellular signaling network, scMLnet utilizes Fisher's exact test [7] to compute the activity score of a specified transcription factor in a given cell type [22]. First, known associations between receptors and transcription factors and between transcription factors and target genes are curated from existing public databases. Next, three sets are constructed: the set of significantly expressed target genes, the set of all possible target genes, and the set of target genes for a given transcription factor. Using the sizes of the intersection sets between these three constructed gene sets, Fisher's test is used to compute the p-value, which represents the activity of a specific transcription factor in a certain cell type. This p-value of a specific transcription factor is lower if there are more highly expressed target genes regulated by it, indicating a higher activity of the transcription factor of interest. Cheng et al. use scMLnet to only consider transcription factors with a calculated p-value lower than 0.05 to be activated in the receiver cells, which are then included in the subnetwork describing interactions from transcription factors to target genes [21]. Similarly, to find links between receptors and transcription factors, Cheng et al. also use Fisher's test is used to compute the activity of receptors according to the calculated p-value. Aside from scMLnet, other tools that use Fisher's test to determine the intracellular interactions are CCCExplorer [23] and scSeqComm [8].

Differential expression-based: Another method, CellCall [112], uses gene set enrichment analysis to calculate an activity score of transcription factors that are activated downstream of a ligand-receptor interaction. CellCall first constructs regulons consisting of a transcription factor and its set of coexpressed target genes. The Spearman's rank correlation coefficient is used to determine gene coexpression between a target gene and a transcription factor; significantly coexpressed genes are retained in the considered regulon. Next, the activity score of the transcription factor is calculated using the gene set enrichment analysis enrichment score for the regulon. If there are multiple transcription factors downstream of a ligand-receptor interaction, then the transcription factor activity score is calculated as the weighted sum of all transcription factors downstream of the interaction.

Network-based: Network-based methods use network analysis methods that utilize the structure of the network to infer likely interactions. scSeqComm uses this type of method to build the receptor-TF subnetwork, first by using existing databases. Next, the PageRank algorithm, a network analysis method in machine learning, is used to calculate the strength of associations between receptors and transcription factors, where the receptor as the seed node. The PageRank scores of each transcription factor is used to measure the association of the transcription factor to a given receptor.

Coexpression-based: Rather than consider interactions from receptors to transcription factors and from transcription factors to target genes, one intracellular method, CytoTalk, constructs an intracellular gene coexpression network between all possible pairs of genes within cells using mutual information [50].

Ligand-target gene network CCC methods

Most CCC methods that consider intracellular signaling networks model intercellular CCC and intracellular CCC separately. In contrast, NicheNet constructs a ligand-target network that connects signaling ligands to downstream target, using PageRank, a network-based method [14]. NicheNet uses ligand-receptor networks, signaling networks, and gene regulatory networks constructed using multiple existing databases. These networks contain interactions from ligands to downstream transcription factors, and gene regulatory interactions between transcription factors and target genes. The data sources are integrated to build a weighted ligand-signaling network and gene regulatory network. Based on the ligand-signaling network, for target genes and upstream ligands of interest, a signaling importance score for each ligand to each target gene is calculated using a Personalized PageRank algorithm. Accordingly, a ligand-gene signaling importance matrix is obtained, where each entry denotes the signaling importance of each upstream ligand to each downstream gene. Next, the final ligand-target gene interaction network is obtained by multiplying the ligand-gene

signaling importance matrix and the weighted integrated gene regulatory network matrix. The entries in this matrix denote thus the regulatory potential of a ligand to a downstream target gene.

Spatial CCC

Spatial transcriptomic technologies are rapidly expanding, which include immunofluorescence-based methods, mass spectrometry-assisted methods, and barcoding-based methods [91]. In contrast to scRNA-seq data in which the spatial information is destroyed, spatial data preserves not only cell-cell heterogeneity information but also spatial positions. Since CCC can only occur between spatially proximal cells, knowing the cellular positions in space allows one to constrain the prediction of potential ligand-receptor interactions and significantly reduce the prediction of false positive interactions. There are two assumptions that are often used when inferring CCC from spatial data. First, CCC is a result of ligand-receptor co-occurrence. Second, the gene expression of cells also depends on their interactions with neighboring cells. Recent advancement in spatial transcriptomics makes it possible to detect genetic information at multiple cells, single-cell and even subcellular resolutions [98]. These advancements makes it possible to explore the above assumptions. Computational tools incorporating spatial information are developed based on these two different assumptions to capture CCC activity. Tools based on the first assumption include Giotto, SpaOTsc, GCN, and DeepLinc, and another tool based on the second assumption is SVCA. In contrast to most tools developed to analyze spatial data directly, SpaOTsc tries to map non-spatial scRNA-seq, which has single-cell resolution and typically higher gene coverage, to the positions of spatial data, and then analyze CCC using the mapped scRNA-data. We will briefly describe existing CCC methods utilizing spatial information below.

Giotto [34] uses spatial information to constrain possible cell interactions and then models CCC strength by calculating a sum-based coexpression score of ligands and receptors. Giotto

first determines whether two cell types are preferentially located in a spatially proximal manner and then tries to identify which ligand-receptor pairs interact between two spatially proximal cell types. For each possible ligand-receptor pair in sender cell type and receiver cell type, the interaction score is a sum-based coexpression score, which is calculated by the weighted average gene expression of ligand and receptor in interacting sender and receiver cells, or in the subset of sender and receiver cells that are spatially close. Giotto then uses a permutation test to assess whether the calculated interaction score is statistically significant. Rather than randomly permuting cell type labels, Giotto randomly permutes the cell locations within the same cell type before calculating the corresponding p-value.

SpaOTsc is a tool that uses spatial data as a reference to provide spatial information to a non-spatial scRNA-seq dataset, from which ligand-receptor interactions are inferred using optimal transport [17]. First, SpaOTsc constructs a spatial metric for the non-spatial dataset. Using the pairwise gene expression similarity between cells in the non-spatial scRNA-seq dataset and cells in the reference spatial dataset, SpaOTsc assigns a position (as a probability distribution over all spatial positions) to each cell in scRNA-seq dataset using optimal transport. Based on this position assignment, SpaOTsc calculates a spatial metric between each pair of cells in scRNA-seq dataset. Then the genes in scRNA-seq data can be viewed as distributions over all cells in the dataset. For a given ligand-receptor pair, SpaOTsc calculates an optimal transport plan from ligand distributions to receptor distributions, where the cost function is defined as the spatial distances between cells. This transport plan gives the ligand-receptor interaction between each cell pair. SpaOTsc then summarizes the interactions between cells across cell type groups to calculate interactions at the level of cell types.

COMMOT is an extension of the optimal transport framework of SpaOTsc that further considers the competition between multiple ligands and multiple receptors [18]. Instead of viewing CCC between cells as an optimal transport plan between probability distributions of

a ligand and a receptor, COMMOT considers CCC between cells as a collection of optimal transport plans for all ligands and receptors that can be coupled simultaneously under some spatial constraint. To consider competition between ligands or receptors, COMMOT assumes that a specific ligand or receptor in a specific cell has limited capacity for interactions that depends on ligand or receptor expression in the specific cells. Thus, in a given pair of cells, for a specific ligand that can bind to different receptors, a stronger interaction with one receptor reduces its potential of interaction with another receptor. The CCC between certain ligand-receptor pair in a pair of cells is given by optimizing a collection of transportation plans. Similar to SpaOTsc, COMMOT then aggregates the interactions between cells in clusters to calculate cluster-level CCC.

GCNG is a spatial CCC method based on supervised learning on graph neural networks [110]. The graph neural network takes a cell neighborhood graph as input. A graph Laplacian is then calculated to encode intercellular spatial relationships in the graph structure, and specific ligand-receptor pair expression in each cell are specified as node attributes. The output of the graph neural network is a binary value which represents whether the specific ligand-receptor interaction exists in this graph.

DeepLinc is a spatial method that uses a variational graph autoencoder (VGAE) to infer CCC [60]. A VGAE takes a cell neighborhood graph, where edges connect neighboring cells, and gene expression as features of the nodes in the graph as input. The encoder of VGAE outputs a latent representation for each cell. Next, the decoder of VGAE calculates cell-cell similarity between by calculating the dot product between latent representations of cells to generate the CCC network.

Spatial Variance Component Analysis models CCC based on the assumption that CCC is a cause of cellular variation, and gene expression variance in cells can be modeled as a linear combination of three components: an intrinsic cell state effect, an environmental effect accounting for the position of the cell, and a cell-cell interaction effect [6]. The three variational

effects are then modeled as multivariate Gaussian distributions whose covariance matrices account for similarity in cell intrinsic state, spatial proximity, and cellular neighborhoods. The covariance matrix that account for similarity of cellular neighborhood is the term of interest that can account for cell-cell interactions, and it is determined by fitting a regression model using maximum likelihood estimation for each target gene.

Significance test

Some CCC tools incorporate a significance test, including permutation test and Kendall's rank-correlation coefficient, on the calculated CCC scores to filter out non-significant interactions. One common method used is the permutation test. After calculating a CCC score of a pair of ligand-receptor from sender to receiver cell types, the cell type labels are randomly permuted for a large number of times to generate a null distribution. The p-value of an interaction is calculated as the probability of obtaining a CCC score that is higher than the original inferred CCC score. Only CCC scores with corresponding p-values that are lower than a specified threshold are kept. The procedures are very similar to permutation based (cell type label permutation) intercellular CCC calculations, though in significance tests they are performed with a different purpose, which is to filter out nonsignificant interactions. Permutation tests are used in CellChat [54] and Graeber et al. There are some other tests that returns a p-value to determine differential expression of ligands or receptors, like ANOVA test in PyMINEr [95]. In some methods, the calculated p-values are corrected for multiple testing using Benjamini-Hochberg correction, including scConnect [52], PyMINEr [95], and ICELLNET [72]. In contrast, Kendall's rank-correlation coefficient is used in scMLnet [22] to filter out weakly correlated links from receptors to transcription factors and links from transcription factors to downstream target genes [21]. For receptor to transcription factor links, they use gene expression matrix of receiver cells to compute the Kendall's rank-correlation coefficient between gene expression of receptor and transcription factors, and then filter out

links with a low Kendall's rank-correlation coefficient. The links between transcription factor and target genes are filtered out similarly.

1.2.2 Downstream analysis of CCC

After infering CCC, several methods go further and extract features based on the properties of the CCC networks, including network centrality analysis [54], network diversity analysis [104], gene entropy analysis [50], communication pattern analysis [54], classification of signaling pathways, and comparison between multiple datasets [14]. In spatial methods, downstream analysis include visualization of the signaling direction in space, and identification of differentially expressed genes affected by CCC [17].

CCC networks can be represented as a directed weighted network, where the nodes represent cell groups and the edges represent directed interactions from sender cell groups to inferred receiver cell groups. Therefore, many CCC downstream analysis methods exploit these network-level features of CCC networks, including network centrality analysis, network diversity analysis, functional and structural similarity of CCC networks [54]. Network centrality, especially node centrality, is used to discover important senders, receivers, mediators and influencers in the CCC networks. The in-degree and out-degree measure the total interaction strength received and sent by a cell group, respectively. Flow betweenness measures the implied ability of a cell group to control the interaction flow between different cell groups. Information centrality provides the amount of control on the information flow. Many other popular centrality measures, including hub score, authority score, EigenCentrality, and PageRank can also help identify important cell groups in CCCs. In addition, network diversity analysis can extract node-level, node-to-node-level, and network-level diversity of multiple CCC networks that represent the CCCs for different biological conditions. For example, node degree centrality has been used to analyze COVID-19 datasets, where it

has been found that there is a decreasing trend of cell interactions received by B cells as the severity of COVID-19 increases. Significant signaling pathways present in the CCC networks can be classified according to pairwise functional similarity and structural similarity, where functional similarity quantifies the similarity of major senders and receivers in each signaling network, and structural similarity measures the topological differences between signaling networks. Non-negative matrix factorization is used to identify how sender cells and receiver cells coordinate with each other with respect to the signaling pathways they use. Apart from network analysis, gene entropy analysis measures the amount of uncertainty in the signaling from a specific gene to another. It is based on an inferred gene-gene network, where each node indicates a gene and each edge indicates the interaction strength between the two genes. The signaling entropy of a gene in the network represents the uncertainty in the signaling activity to other genes, and is calculated using Shannon entropy [87].

Comparison of CCC features between multiple datasets is another popular downstream analysis. For example, CellChat implements methods to compare differences in interaction strengths, major sender and receiver cell groups, interaction strengths in signaling pathways, and upregulated and downregulated ligand-receptor pairs between conditions [54]. Network diversity analysis also enables the comparison of differences in CCC diversity across multiple conditions [104].

Among all spatial CCC methods, COMMOT performs some downstream analysis based on the inferred CCC [18]. The spatial vector field of signaling directions of incoming and outgoing cells are visualized in a spatial signaling direction plot. Also, differential gene analysis is performed to identify differentially expressed genes that are affected by the spatial CCC.

1.2.3 Experimental Validation

As many computational tools have been developed to measure CCC levels, it is important to validate the predicted interactions. There are different ways to validate the calculated CCC levels, including experimental methods, spatial colocalization and literature support.

Experimental validation can be performed in three different ways: (1) Validation of the expression of ligands and receptors involved in CCC using proteomics, enzyme-linked immunosorbent assay, immunohistochemistry or western blots. (2) Functional assessment of CCC roles by in vivo or in vitro experiments using activators or inhibitors of involved ligands and receptors. (3) Visualization of CCC between neighboring cells using microscopy and immunostaining, single-molecule fluorescence in situ hybridization, and flow cytometry [4].

Spatial colocalization validation assumes that spatially adjacent cells are more likely to interact than cells that are further apart. Thus, a reasonable CCC score should have the property that it is larger between adjacent cells and lower in further apart cells. Since this type of validation requires spatial information of cells, it is normally performed on spatial transcriptomics dataset.

Evidence from biological literature is also a popular method to indirectly validate the predicted CCCs. For example, when studying CCC between dermal condensate (DC) and epithelial placode cells in hair follicle morphogenesis, CellChat inferred that epithelial placode cells distinctly secrete Fgf20 ligand to all DC states, which is consistent with the previously known role of placode-derived Fgf20 signaling.

1.2.4 Analysis of CCC in cancer

Cancer is a current major health concern worldwide. Genetic and epigenetic alterations induce changes in cell-cell interactions that let cells escape homeostatic controls and drive cancer. Changes in CCC influence not only cancer cells but also other important components in the tumor microenvironment, including immune cells, endothelial cells, surrounding stromal cells, and others. Ligand-receptor interactions between cancer cells and normal cells and intracellular signaling are capable of regulating multiple crucial cellular processes during the progression of cancer, including cell growth and division, cell apoptosis, cell motility and invasion, angiogenesis, inflammation and immune suppression [85, 46, 45]. For example, Sever and Brugge give a comprehensive examination of how changes in PI3K-Akt and Ras-ERK interactions to can cause different characteristic features of cancer to emerge [85]. Due to its clear importance in cancer, the analysis of CCC can potentially greatly enhance our understanding of cancer mechanisms. In pharmacology, targeting the interacting cells or mediators of cell interactions has been proven to be effective in multiple tumor treatments [23, 85].

Using computational CCC tools, one can quantitatively study the upregulated or downregulated ligands and receptors within and between normal cells and tumor cells to potentially identify novel cell interactions, raise new hypothesis and provide evidence on cell interaction mechanisms in cancer development. Here, we highlight but a few applications. For example, Choi et al. used CCCExplorer [23] to identify new ligand-receptor interactions, including IL6-IL6R and WNT11-FZD7, between macrophages and tumor cells in non-small cell lung cancer (NSCLC). Choi et al. also verified the presence of AREG-EGFR signaling from monocytes to tumor cells, which has been reported in many other cancer types. Since AREG has been reported to inhibit apoptosis, the presence of this interaction suggests a potential tumor-promoting role of monocytes in NSCLC (Choi et al. 2015). In another study, using human scRNA-seq data of head and neck squamous cell carcinoma (HNSCC) tumors,

Browaeys et al. used NicheNet to investigate the hypothesis that cancer-associated fibroblasts (CAFs) may regulate genes involved in the partial epithelial-to-mesenchymal transition (p-EMT) program, as originally proposed by Puram et al. (Browaeys, Saelens, and Saeys 2020). They inferred the top interactions between CAF-ligands and p-EMT target genes in cancer cells, including interactions from TGFB3 to TGFBI, LAMC2, and TNC, shedding further light on the potential regulatory roles of CAF-ligands on p-EMT target genes. Arnol et al. [6], the developers of SVCA, studied a breast cancer Imgaging Mass Cytometry (IMC) dataset. Using SVCA, they inferred that CCC components can explain up to 25% of the gene expression variance, where marker genes of immune cells, including CD44, CD20, CD3 and CD68, comprised the largest CCC effects. By using CellCall to study CCC between immune cells from six types of cancer, including liver hepatocellular carcinoma, non-Hodgkin lymphoma, non-small cell lung cancer, kidney renal clear cell carcinoma, colorectal cancer, and breast invasive carcinoma, Zhang et al. inferred that monocytes and macrophages receive the largest amount of signals from other immune cell types, indicating their roles as dominant signaling receivers among all immune cell types in multiple cancers [112].

Due to frequent mutations occurring in cancer, there are some limitations of using computational CCC tools to infer cell interactions in cancer. First, cancer-causing mutations may lead to distortions in inferred interactions. For example, a recent study has shown that in melanomas, lung and colon cancers, the gene BRAF mutation V600E will induce a novel interaction between BRAF and KEAP1 [68]. Such distorted interactions caused by mutations may not be captured by currently available CCC methods or databases. Furthermore, for methods that consider downstream effects of ligand-receptor interactions, if one of the downstream genes is mutated, this could affect the entire downstream network significantly, which may, in turn, affect CCC inference.

1.2.5 Future Directions

With the vastly increasing number of scRNA-seq and spatial datasets and the fastly developing technologies, opportunities arise as we try to increase our understanding of CCC and improve the reliability of CCC tools.

1. Incorporating spatial information Incorporating spatial information has many advantages. First, cells primarily interact with each other locally and across limited spatial distances. While scRNA-seq provides the gene expression of ligands and receptors in cells, which is important for the detection of CCC, it does not capture the physical positions of cells, missing an important spatial constraint on possible interactions. Second, non-spatial data cannot accurately infer CCC that occur through physical cell-cell contact. It is possible that future work can model the cell-cell contact by incorporating cellular spatial information. Third, most current CCC methods calculate the interaction strengths between groups of cells to reduce false positive estimates. However, these methods implicitly assume that all cells within the sender and receiver groups are in signaling range. It may be the case that only a subset of these cells are close enough to interact. Thus, it is crucial to develop more tools that can map a spatial distance between cells, by either equipping spatial positions to non-spatial cells, or by using ST to profile the cells together with their spatial positions.

2. Incorporating temporal information Most CCC methods rely on the average ligand and receptor expression in different groups of cells to infer CCC from scRNA-seq data. However, as scRNA-seq is a static snapshot of gene expression, these methods an important aspect of CCC, which is that it is dynamic. Incorporating time information provides a new angle of studying CCC in general. Apart from learning how CCC evolves with respect to time, time information may also aid in finding associations between CCC and key biological processes that may be important in cancer, including cell differentiation, organ development, disease progression, and immune response. To account for temporal information, one could analyze

CCC across a time series dataset of scRNA-seq data sampled from the same biological system at multiple time points, or by applying a pseudotime ordering method on a single dataset and analyzing CCC across inferred pseudotime branches [60, 54].

3. Competition and cooperation between CCC Currently, most CCC tools rely on the assumption that ligand-receptor interactions occur are independently, which may not be the case. For example, researches have shown that the invasion of a human mammary tumor is driven by both paracrine signaling to and from host macrophages and autocrine signaling between tumor cells. It may therefore be the case that these interactions co-interact. Furthermore, multiple ligands compete to bind to the same set of receptors, but current methods assume that these ligands can bind to the same receptors across all possible receiver cells. Thus, future work could incorporate the dependency between ligand-receptor interactions to improve accuracy of inferred CCC. There has been some recent research in this direction [109].

4. Multi-omics integration The overwhelming majority of current CCC methods focus on intercellular signaling and do not account for the intracellular downstream response (with some exceptions). A key aspect of intracellular signaling is the activation or in activation of transcription factors after ligand-receptor binding. One example is the activation of transcription factors in the SMAD family due to TGFb signaling. Once a transcription factor is activated, it binds to specific regions of non-coding DNA to modulate the transcription downstream target genes as a response to cell signaling. Given their evidently important role in the intracellular signaling network, accounting for the signaling gene regulatory network facilitated by transcription factors can provide extra information when studying CCC. Such information can be obtained through several recently developed technologies. For example, advances in mass spectrometry have enabled the high-throughput measurements of post-translational modulation of signaling proteins. Other technologies like single-cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) and Chromatin

immunoprecipitation followed by sequencing (ChiP-seq) provide information on chromatin structure and accessibility, which regulate binding affinity of transcription factors and their consequent regulation of downstream target genes [] (Yan et al. 2020). The incorporation of measurement of transcription factor activity from other omics data is promising for both the inference and validation of CCC. Furthermore, as ligand-receptor interactions occur at the protein level, gene expression information alone cannot measure CCC directly. It is thus potentially promising to incorporate other technologies that provide protein level information. For example, cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) can integrate protein and transcriptome measurements into a single-cell readout.

1.3 Discussion

The development of both scRNA-seq and ST technologies has generated a new, rapidly expanding field of CCC studies. A wide range of computational tools have been developed recently to model and infer both intercellular and intracellular interactions in different species (e.g. human or mouse), and to analyze the downstream features of these constructed CCC networks. The currently available tools use different computational methods to model CCC from gene expression, based on their modeling assumptions. Even though ground truth datasets are currently lacking, some efforts have been made to compare between different CCC methods. For example, one study compares seven CCC methods methods with respect to: agreement of inferred CCCs with spatial colocalization, cytokine activities, and receptor protein abundance, concluding that the predicted CCC are coherent with these modalities in general [32]. They also found that different ligand-receptor databases used by different CCC methods have different biases towards certain pathways and tissue-enriched proteins. Another comparative study evaluates 16 CCC methods by integrating scRNA-seq with ST data, concluding that CellChat, CellPhoneDB, NicheNet, and ICELLNET show better per-

formance in alignment with spatial information and in scalability. For future work, we envision that CCC methods will be benchmarked based on their agreement with existing biological evidence. Though a large number of single-cell-focused tools have been developed, there are still limitations of current CCC inference methods. First, scRNA-seq data contain biological and technical noise in experimental processing [24], which may obscure the true gene expression of signaling genes. Second, ligand-receptor interactions occur at the protein level, meaning that the gene expression measurements provided by RNA-based measurements only give an indirect measurement of ligand and receptor expression. Third, most methods prioritize ligand-receptor interactions, but these are by no means the only form of CCC. For example, a recent method, NeuronChat [113] infers CCCs due to neurotransmitters consisting of groups of small molecules that are typically excluded from current ligand-receptor databases and corresponding receptors. Such interactions are not accounted for by other CCC methods and are known to play a role in glioma progression [42, 55]. Fourth, while experimental validation, spatial colocalization, and literature support provide validation of the computed CCC scores from different tools to some extent, there is still a lack of ground truth datasets for which the underlying CCC networks are truly known, preventing substantial benchmarking of CCC inference methods. We propose that future work should try to improve the following aspects in developing computational tools to study CCC: (1) incorporate spatial information to improve the accuracy of calculated ligand-receptor scores; (2) incorporate temporal information to understand the dynamic process of CCC changes in time; (3) take into account the dependency of ligand-receptor interaction; (4) integrate scRNA-seq with other omics technologies to gain information at the protein and epigenetic levels and improve modeling and inference of CCC. With these points in mind, it is clear that the use of single-cell technology to analyze CCC will continue to grow at a rapid pace in the following decades to come, and so too will the insights that come with this growth.

Chapter 2

Analyzing Diversity of Cell-cell Communication Networks via Single-cell Transcriptomics

2.1 Introduction

Cell–cell interactions (CCC) manipulate multiple biological processes, including organismal development, homeostasis, and immune responses [4]. When cells do not interact properly, disease occurs. The recent advances in single-cell RNA-sequencing (scRNA-seq) offer great opportunities to decipher CCC through coordinated gene expression of ligand–receptor pairs. Using a diverse range of strategies, many tools have been designed to infer CCC from scRNA-seq [4, 3]. These computational tools can be grouped into four categories: differential combination–based, network-based, permutation-based, and array-based. The majority of CCC inference methods are designed to analyze CCC for a single condition. As characterizing differences in CCC activity across multiple conditions—for example, between healthy

and diseased tissue—helps elucidate the diverse mechanisms of CCC-mediated responses to disease, more so than analyzing condition-specific CCC networks in isolation, recent efforts have focused on developing methodologies to systematically study the differences in CCC between multiple conditions [108, 41, 103]. However, the majority of previous CCC analyses that analyze differences between conditions tend to analyze CCC networks inferred from a single scRNA-seq sample from each condition or analyze CCC networks constructed by aggregating several samples. These analyses result in deterministic characterizations of CCC and fail to take into account the underlying variability across the groups of samples that constitute each biological condition.

As single-cell omics field is expanding, data are becoming more readily accessible. Advances in sequencing technologies have enabled more sample replicates to be obtained at cheaper prices and a greater number of cells that can be sequenced per replicate [90]. While including multiple samples can obscure biological variation via batch effects [64], having multiple samples per condition enables one to characterize the variability of key aspects such as CCC at the "group level", where multiple samples comprise a single condition. Being able to account for variability across samples, such as those sampled from human patients, is particularly important for diseases where host response is associated with severity of disease progression, such as COVID-19 [73]. Indeed, there are a number of scRNA-seq studies of COVID-19 where multiple samples have been obtained for both healthy and diseased status [62, 111, 66, 77, 88, 57]. In particular, Ren et al. (2021) produced scRNA-seq data containing 1.5 million cells from 196 patients across five COVID-19 conditions. As more large-cohort scRNA-seq studies become available, it is important to develop methodologies that account for and analyze the sample-to-sample variability within and across biological conditions.

Thus, in this article, we propose a study on diversities of three classes of features of CCC networks within groups of different conditions. More specifically, since a CCC network can be represented as a directed weighted graph, where the nodes correspond to cell states and

edge weights correspond to interaction strengths, we can extract CCC features at three levels: single node, node-to-node, and whole graph. We apply these methods to a large single-cell transcriptomic COVID-19 dataset comprising peripheral blood mononuclear cells (PBMC) sampled from 181 patients and analyze CCC diversity patterns under different COVID-19 conditions. By analyzing each specific CCC feature and comparing its diversity across different sample groups comprising biological conditions, we are able to observe consistent and biologically meaningful patterns aligned with the progression and subsequent convalescence of COVID-19.

2.2 Results

2.2.1 Analysis of CCC in different COVID-19 Conditions

To study CCC differences across COVID-19 patients under different disease conditions, such as healthy, moderate disease, severe disease, convalescent from moderate disease, and convalescent from severe disease, we used CellChat and studied the interactions among all 3 cell types: myeloid, B cells, and T cells.

CCC between any two cell types consist of interactions via different signaling pathways. First, we compared the CCCs of each cell pair in five groups (Figure 2.1a). We observed that autocrine signaling of T cells only exists in moderate and severe groups. Myeloid autocrine signaling decreased in severe samples, compared to healthy controls and moderate samples. In samples in the convalescence stage from severe disease, myeloid autocrine was recovered compared to the severe group. Compared to healthy controls and moderate patients, signaling from T cells to B cells was decreased in the rest of the condition groups. Interestingly, interaction strengths from T cells to myeloid cells were consistent over all the five groups. We then examined the differences in total interaction strengths in the five groups

(Figure 2.1b) and found that the interaction strengths were higher in the control and moderate groups and the lowest in the severe group. We also compared the incoming and outgoing interaction strengths of different cell types (Figure 2.1c). The total interaction strength of B cells as a receiver is always stronger than as a sender, while the total interaction strength of T cells as a sender is always stronger than as a receiver. Myeloid cells interact similarly in strengths as senders and receivers.

Having obtained a general knowledge on CCC strengths across five conditions, we then investigated upregulated and downregulated signaling in each group compared with the remaining four groups (Figure 2.1d). CellChat predicted significant upregulation of three ligand–receptor pairs in control patients, consisting of the ligands TNFSF13, IL16, and TNFSF13B, which were sent by B cells and received by B cells; significant downregulation of ligand–receptor in control patients, consisting of the ligands CCL3 and TNFSF13B, which are sent from myeloid cells to B cells and from T cells to B cells. We also observed five upregulated ligand–receptor pairs in severe patients, with the corresponding ligands being IL1B, ANXA1, CCL3, TNFSF13B, and MIF, and two dowregulated ligand–receptor pairs, where the ligands are IL16 and LGALS9. Interestingly, there were no upregulated ligand–receptor pairs in the severe convalescence group.

Next, we examined functional and structural similarities of different signaling pathways in five disease conditions (Figures 2.1ef). We observed that most pathways were grouped together under both similarity measures even if they were of different disease conditions.

To test robustness of CCC output and the conclusions from CellChat, we downsampled the original data to 80% of each original condition groups. We found that the CCC networks were robust to downsampling (Figure 2.1a, Figure A.1), as had been found previously in [54]. The majority of conclusions from CellChat hold under downsampling. However, the upregulated and downregulated ligands change (Figure 2.1d, Figure A.1c). Specifically, the upregulated genes all disappear in the downsampled version, and downregulated ligands in

moderate and convalescence from severe group are different.

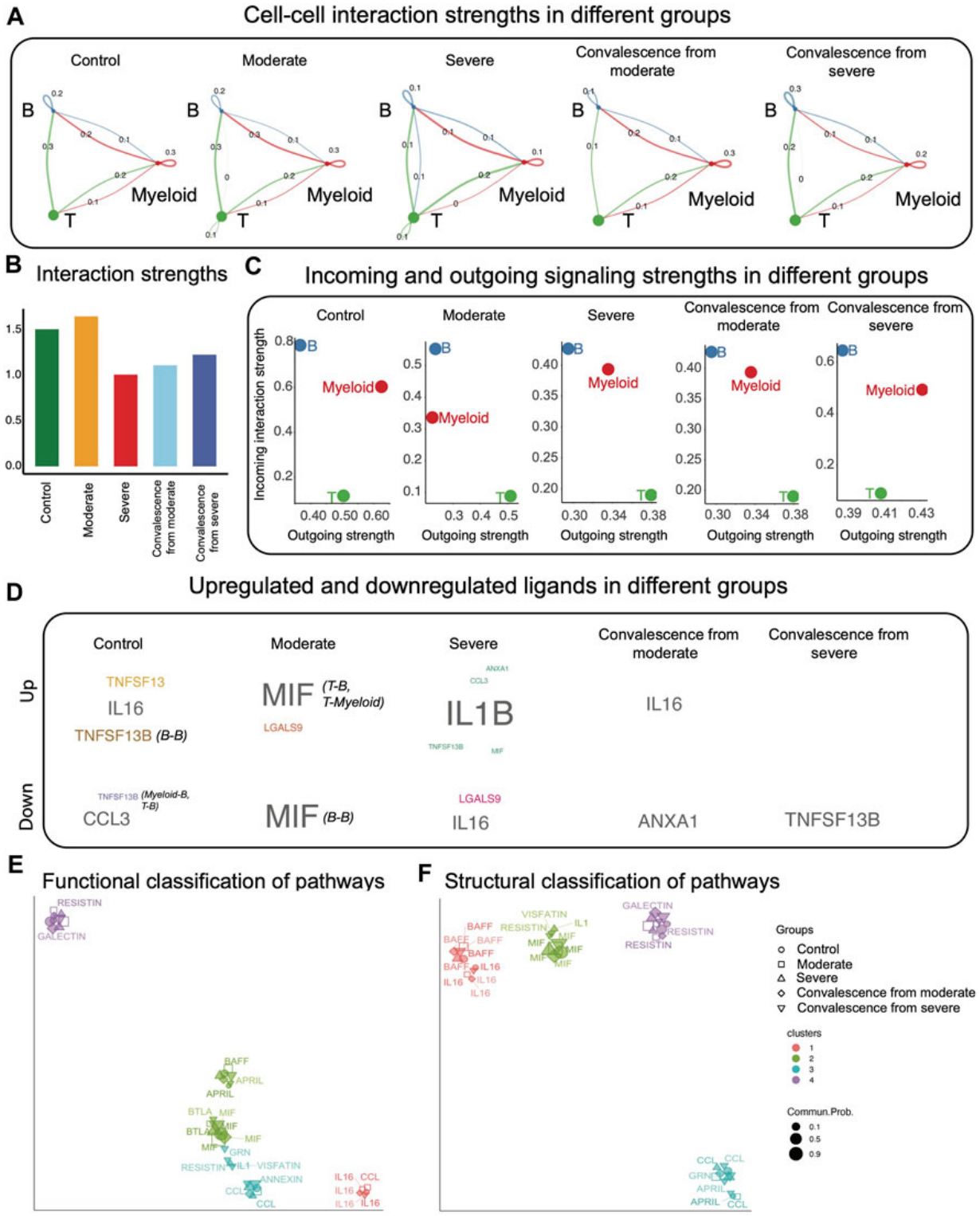


Figure 2.1: **CCC analysis using CellChat.** (a) CCC strengths. See next page for full caption.

Figure 2.1 (continued). (b) Total interaction strength in different groups. Control and moderate groups have higher total interaction strength. (c) Total incoming and outgoing signaling strengths in different groups. B cells interact stronger as receivers than as senders, while T cells interact stronger as senders than as receivers. (d) Upregulated and down-regulated ligands in different groups. (e) UMAP projection of signaling pathways based on functional similarity. (f) UMAP projection of signaling pathways based on structural similarity.

2.2.2 Analysis of CCC in MIF pathway

The log-scaled information flow calculated by CellChat is shown in Figure 2.2a. We observed that the MIF pathway contains the highest total interactions over all signaling pathways, while other pathways contain little amount of interactions. We then calculated the percentage of patients in different groups using each pathway (Figure 2.2b). Out of all pathways, we found that the GALECTIN, MIF, ANNEXIN, BAFF, and IL16 pathways were present in more than 60% of patients, while other inferred pathways were present in fewer than 60% of patients in each condition. To find the most differential pathway across conditions among these popular pathways, we calculated pairwise Canberra distance of CCC in these pathways between each condition and took average among all pairs. We observed that the MIF pathway is also the most differential pathway across conditions (Figure 2.2c). Furthermore, as recent studies have shown that there is a strong correlation between the serum level of MIF and the severity of COVID-19, suggesting that the MIF serum level may be a useful predictor of COVID-19 disease severity [2, 9, 31], we focused our analysis on CCC diversities in the MIF pathway for the remainder of this study.

We observed that in the MIF pathway, autocrine signaling of T cells only exists in two diseased groups, which indicates that most T cells' autocrine signaling in total CCC comes from the MIF pathway (Figure 2.2d). We found that unlike the rest of the groups, myeloid cells exhibit no autocrine signaling in severe and convalescence from severe groups. We also observed an interaction from B cells to T cells occurring in the severe group, which does not

exist in the rest of the groups. We found that in the MIF pathway, the interaction from T cells to B cells was the highest interaction, except for convalescence from severe group in which it is the second highest. This indicates that T to B signaling functions as the major interaction in the MIF pathway. The most active ligand–receptor pairs in the MIF pathway are then compared across conditions (Figure 2.2e). In the control, moderate, and convalescence from severe conditions, the most active multi-receptor unit is the heteromeric complex containing CD74 and CD44. However, in the severe and convalescence from moderate conditions, the most active multi-receptor unit is the heteromeric complex containing CD74 and CXCR4, indicating differences in the types of MIF-specific interactions across COVID-19 conditions. Next, applying the network centrality analysis of the MIF signaling network shows that in severe and convalescence from severe groups, myeloid cells lose their function as a mediator (Figure 2.2f). This indicates that in those two groups, the myeloid cells have a diminished role as a gatekeeper of CCC.

To validate the robustness of these results, we repeated the analysis of CCC after downsampling the original data to 80% of the sample sizes for each condition group. We observed that the CCC networks and network centrality conclusions of MIF pathways are robust to downsampling (Figure 2.2, Figure A.2).

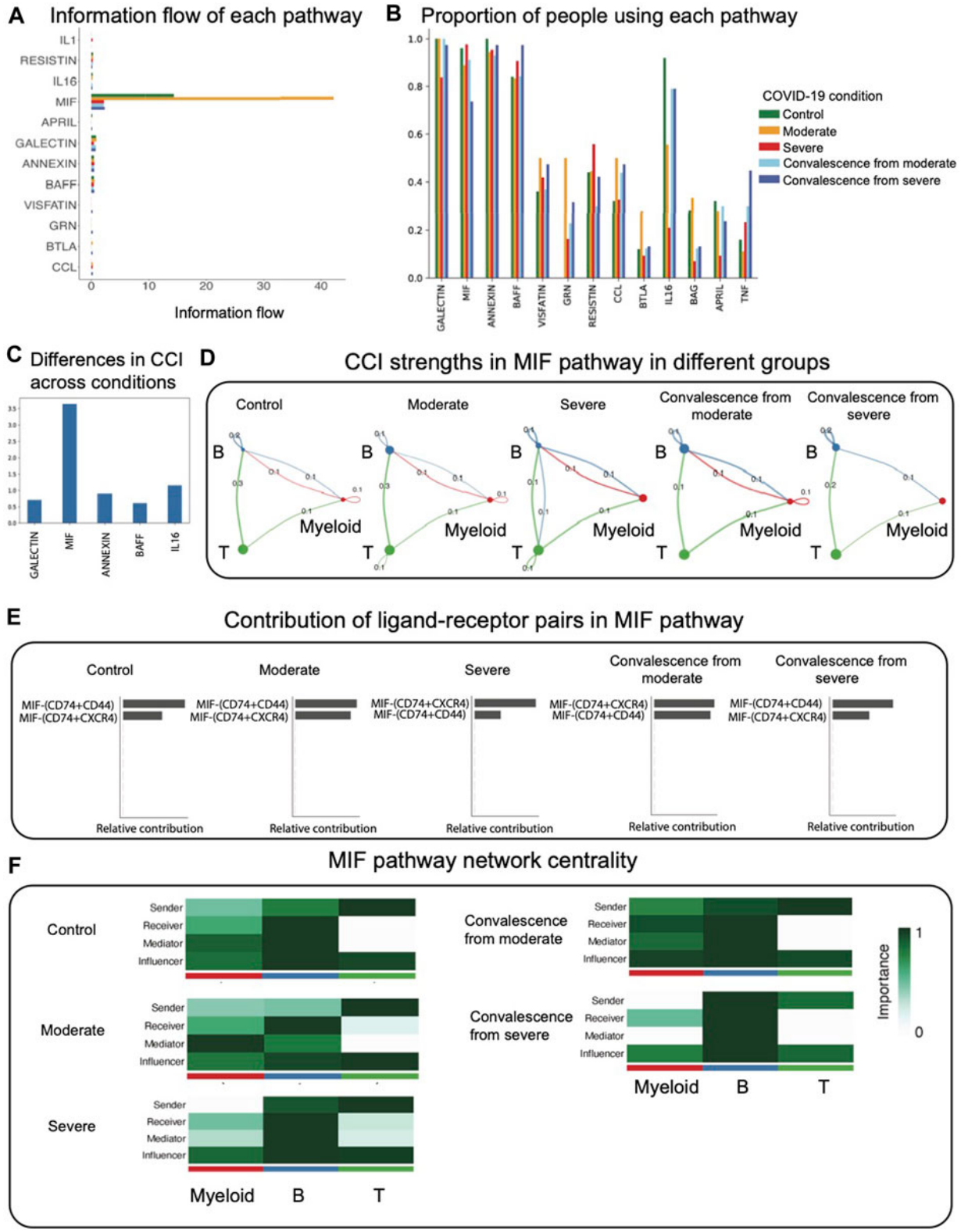


Figure 2.2: **CCC analysis using CellChat focused on the MIF pathway.** (a) Information flow (total interaction strengths) of each pathway. The MIF pathway contains the most amount of information. See next page for full caption.

Figure 2.2 (continued). (b) Proportion of people using each pathway in different groups. The majority of CCC occurs through GALECTIN, MIF, ANNEXIN, and BAFF. (c) Average pairwise Canberra distance of CCC across conditions in the popular pathways. (d) CCC strengths in the MIF pathway. Each edge represents an interaction; the color of the edge matches with the sending cell type. (e) Relative contribution of ligand–receptor pairs in the MIF pathway. (f) Network centrality scores of the MIF pathway. In severe and convalescence-from-severe groups, myeloid cells lose their function as mediators.

2.2.3 Cell type diversity of CCC in MIF pathway

We applied three cell type diversity statistics of CCCs, including degree, diversity, and entropy, to study the diversity in interactions of each cell type across five groups in COVID-19.

We first calculated the cell type diversity to measure the number of different cell types with which sender or receiver interacts (Figure 2.3c-d). A higher cell type diversity means the cell type interacts with more cell types, indicating a higher variability of interactions. We observed that T cells are the cell type with the highest mean cell type diversity as receivers but the lowest cell type diversity as senders. When comparing between groups of conditions, when myeloid cells or B cells are senders, we found a clear decrease in mean diversity from control to moderate to severe groups. Also, when these two cell types are senders, mean diversity decreases from the moderate convalescence group to the severe convalescence group. These findings indicate that as disease progresses, myeloid and B cells as senders tend to interact with fewer receivers. When T cells are sender cells, there is an increase in diversity from control to moderate to severe groups and an increase from the moderate convalescence group to the severe convalescence group. This means that unlike myeloid and B cells, T cells tend to interact with more cell types in severe patients. Interestingly, when T cells are sender cells, we observed that the mean of cell type diversity in moderate and severe convalescence groups are less than moderate and severe groups respectively, which indicates a loss of interaction diversity in convalescent stages. Looking at diversity of receiver cell

types, we observed a decreasing pattern of diversity of myeloid as receiver cells from control and moderate groups to severe groups, and the diversity in two convalescence groups are less than their corresponding groups in disease. This indicates that receiver diversity decreased in the severe group and the convalescence groups. We also observed that when B or T cells are receiver cell types, the variation of diversity is the highest in the convalescence from severe group.

We then calculated the cell type entropy to measure whether a sender interacts evenly with different receivers in strengths (Figures 2.3e-f). A high cell type entropy means the cell type interacts evenly with different cell types, while a lower entropy means the interactions of a sender are concentrated on few receivers. Comparing the cell type entropy across different COVID-19 conditions, we observed that both diversity and entropy had a similar trend. For example, the decreases in diversity in myeloid and B cells as senders from control to moderate to severe also occur in entropy. This means that a sender (receiver) in a typical sample will not have extremely uneven interactions with its receivers (senders). Otherwise, if myeloid cells are communicating with all three cell types but the interactions are concentrated towards T cells, then its entropy will be close to minimum but diversity will be three, which is maximum, causing the two measures to be significantly different.

As a standard measure of interaction strength, the in-degree and out-degree for all nodes (cell types) across all five condition groups are also calculated. The in-degree of a node is the total amount of interactions a receptor receives from all senders. Similarly, out-degree is the total amount of interactions a sender sends to all receivers. Interestingly, the trends we observed in cell type diversity and entropy were similar in senders but not so in receivers. When T cells are receiver cells, we observed a decrease from control to moderate to severe groups and from moderate convalescence to severe convalescence groups. But the consistent decreasing pattern from control to moderate to severe groups does not occur in receiver diversity nor receiver entropy of T cells. When B cells are receiver cells, there is a consistent decrease from control all the way to severe convalescence groups, which does not occur in

the previous measures either. The aforementioned analysis implies that the three cell type diversity statistics can be used together to reveal a comprehensive picture of its interactions from three angles: interaction strength, number of senders or receivers the cell type interacts with, and the evenness of distribution of interaction with different senders or receivers. We applied these three cell type diversity metrics on a COVID-19 dataset to summarize the overall diversity trend of cell types in different disease conditions.

2.2.4 A random walk inspired node-to-node information flow in CCC graphs in MIF pathway

To study the total amount of information flow among different nodes in CCC graphs, we calculated the information flow proximity matrix of each patient in the COVID-19 dataset. Comparing the distributions of node-to-node or cell-type-to-cell-type information flow across the five groups of different disease conditions, we observed several patterns (Figure 2.4a). When myeloid cells are sender cells, there is a decrease in mean information flow from control to moderate to severe groups and from moderate convalescence to severe convalescence groups. This observation indicates that the importance of myeloid cells as sender cells in a network decreases in these groups. The mean information flow from B cells to each cell type is lower in severe patients than in healthy and moderate COVID-19 patients and is lower in severe convalescence patients than in moderate convalescence patients. We also made an interesting observation that when B is the sender cell type, in each of five groups of conditions, mean and variance of information flow increases from myeloid to B cells and from myeloid cells to T cells as receivers. We found that T cells mainly function as receivers rather than senders, and there is almost no information sending from T cells to myeloid in all five groups.

A principal component analysis (PCA) is then performed on the information flow proximity matrices of each patient. Patients are then visualized with respect to the first two principal components (Figure 2.4b). While nonlinear projection methods, such as t-distributed

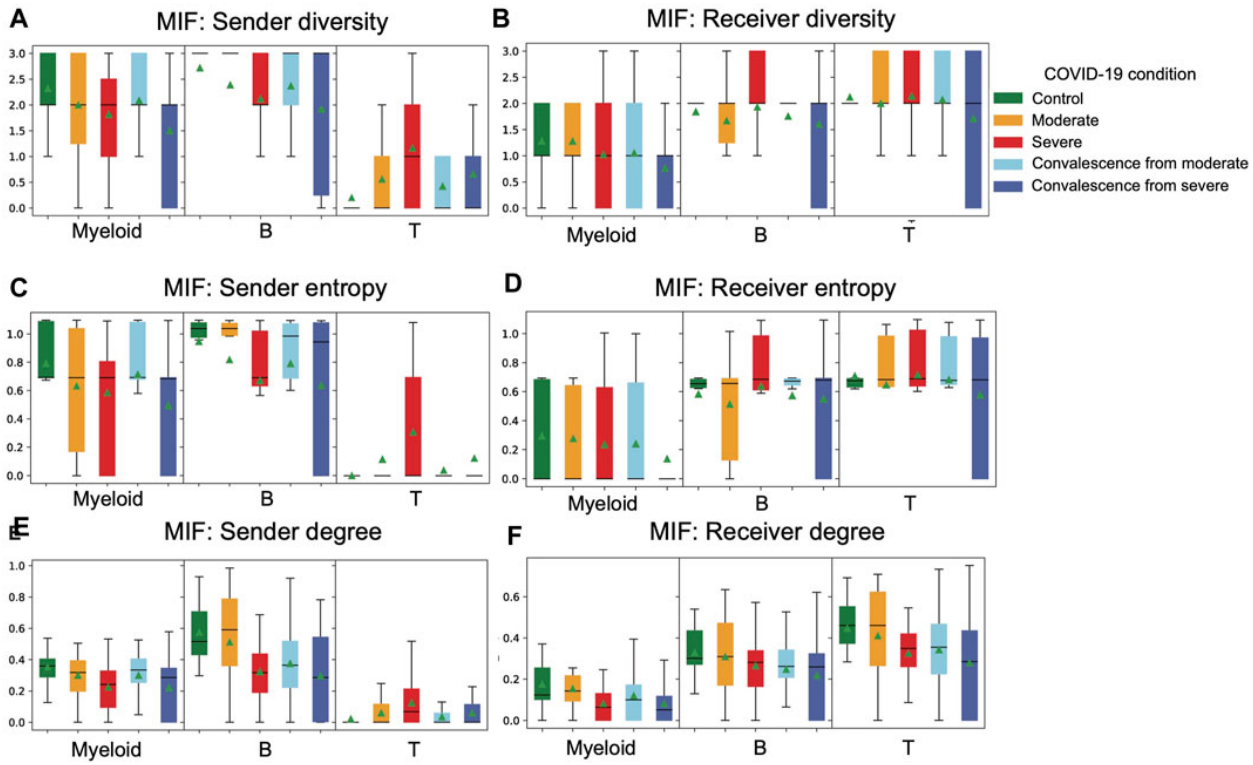


Figure 2.3: **Distribution of cell type diversity statistics in different COVID-19 conditions are visualized.** (a) Distribution of sender diversity in MIF pathway. There is a consistent pattern of decreasing diversity from control to moderate to severe groups and from moderate convalescent to severe convalescent groups in myeloid and B cell types as senders. The pattern in T cell type is different from the patterns in myeloid and B cell types.(b) Distribution of receiver diversity in the MIF pathway. There is a decreasing pattern of diversity from control to moderate to severe to moderate convalescence and to severe convalescence groups in myeloid as a receiver. (c) Distribution of sender entropy in the MIF pathway. The patterns are similar to sender diversity. (d) Distribution of receiver entropy in the MIF pathway. The patterns are similar to receiver diversity. (e) Distribution of sender degree in the MIF pathway. The patterns are similar to sender diversity and sender entropy. (f) Distribution of receiver degree in the MIF pathway. When B cells are receiver cells, there is a consistent decreasing trend of receiver degree from control all the way to severe convalescence groups.

stochastic neighbor embedding (tSNE) or UMAP can be used, we used linear PCA for visualization to maximize interpretability, as tSNE and UMAP require careful calibration of hyperparameters to ensure robust results (Figure A.3). We can see that with respect to information flow, severe patients differentiate from healthy controls and moderate patients.

2.2.5 Canberra node-to-node similarity in CCC graphs in MIF pathway

To study how similar two cell types as senders (receivers) are with respect to their interaction strength with each receiver (sender), we calculated the modified Canberra proximity matrix for each patient and plotted the distribution of modified Canberra similarity of each pair of senders or receivers (Figures 2.4c-d). There are several interesting observations we can make. First, myeloid cells and B cells as senders have the highest similarity among sender pairs in all five groups. Second, myeloid cells and T cells as senders have a similar pattern with B cells and T cells as senders across different groups. These two findings may indicate that for each receiver, myeloid cells and B cells are very similar as senders, but T cells seem to function differently as senders. We also observed that of all possible receiver pairs, B cells and T cells have the highest similarity. Myeloid cells and B cells or myeloid cells and T cells as receivers have similar patterns of similarity across different groups. Similarly, this indicates that for each sender, B cells and T cells as receivers act similarly, but myeloid cells may function differently. We then compared the differences in five different groups of conditions. When comparing myeloid cells and T cells as senders, or B cells and T cells as senders, there is an increase of mean modified Canberra similarity from control to moderate to severe groups and from moderate convalescence to severe convalescence. This indicates that as disease progresses, the relative difference of interactions of myeloid cells and T cells as senders tend to decrease as a response to disease. Similarly for B cells and T cell types, when observing patterns in receiver pairs, we found that severe group has the lowest similarity between myeloid cells and B cells and between myeloid cells and T cells. Thus in severe patients, the

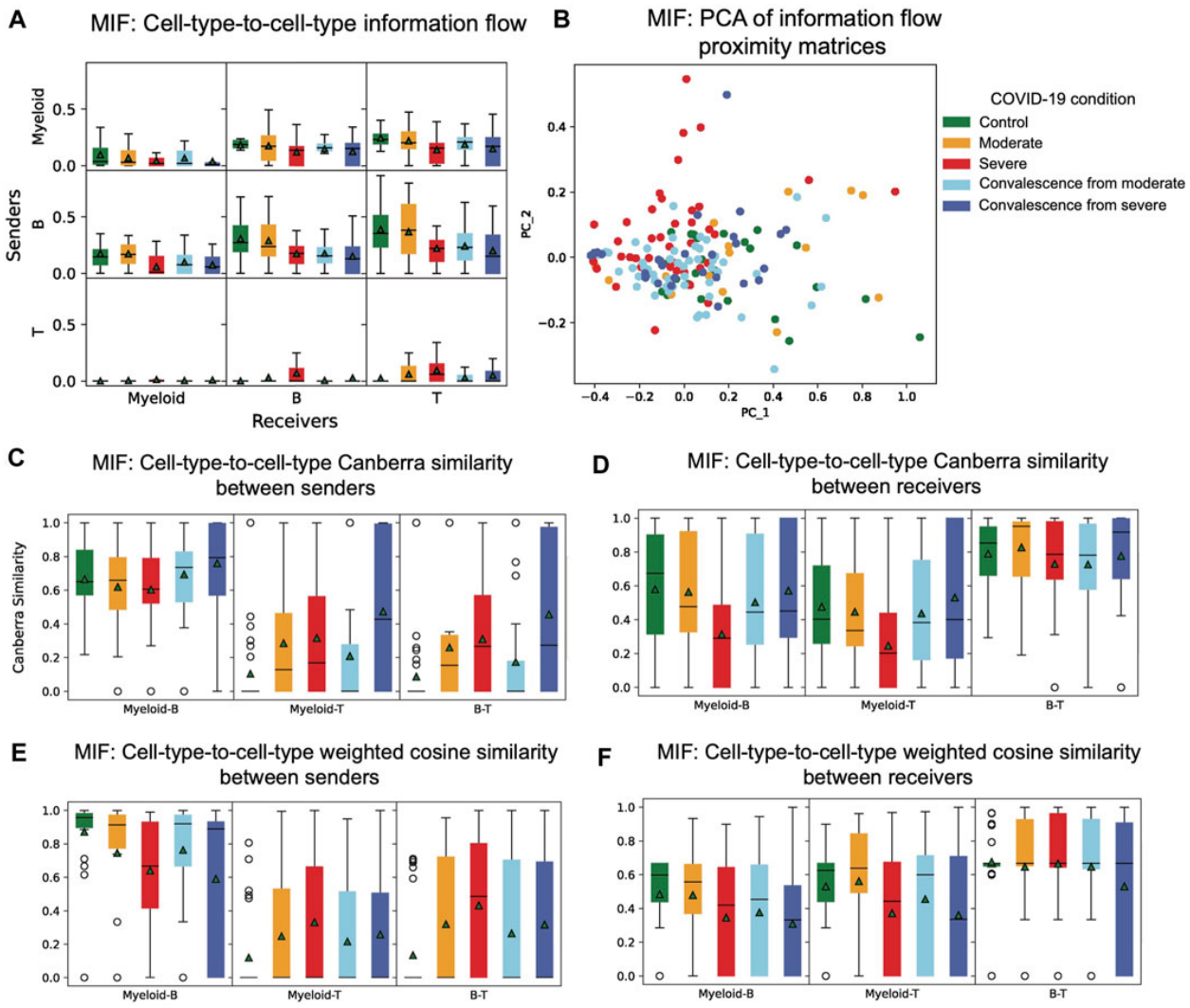


Figure 2.4: **Distribution of cell-type-to-cell-type features.** (a) Distribution of information flow between cell types of five COVID-19 conditions in the MIF pathway. When B cells are the sender cell type in each group of condition, mean and variance of information flow increases from myeloid to B cells to T cells as receiver cell types. (b) PCA of information flow proximity matrices of five COVID-19 conditions in the MIF pathway. There is a clear pattern that differentiates severe patients from healthy controls and moderate patients. (c) Distribution of cell-type-to-cell-type modified Canberra similarity between sender cell type pairs of five COVID-19 conditions in the MIF pathway. Myeloid and B cell types are similar as senders. (d) Distribution of cell-type-to-cell-type modified Canberra similarity between receiver cell type pairs of five COVID-19 conditions in the MIF pathway. B and T cell types are similar as receivers. (e) Distribution of cell-type-to-cell-type weighted cosine similarity between sender cell type pairs of five COVID-19 conditions in the MIF pathway. When comparing myeloid and B as sender pairs, there is a decreasing pattern from control to moderate to severe and from moderate convalescence to severe convalescence groups. On the contrary, when comparing myeloid and T cells or B cells and T cells sender pairs, the patterns are the opposite. (f) Distribution of cell-type-to-cell-type weighted cosine similarity between receiver cell type pairs of five COVID-19 conditions in the MIF pathway. When comparing myeloid and B as receiver pairs, there is a decreasing pattern from control and moderate groups to severe and to moderate convalescence and to severe convalescence groups.

relative difference of interactions between myeloid cells and B cells or between myeloid cells and T cells as receivers decreases.

2.2.6 Weighted cosine node-to-node similarity in CCC graphs in MIF pathway

The weighted cosine similarity between two senders measures if the senders have similar relative composition of ligands. A high similarity between two senders indicates that interactions of both senders with the same receiver are highly positively correlated. We also have a similar interpretation for weighted cosine similarity of two receivers. We calculated weighted cosine proximity matrices of each patient in the dataset and plotted the distribution of similarity of each pair of senders and receivers (Figures 2.4e-f). In comparing myeloid and B as both senders and receivers, we observed a decrease in mean similarity from control to moderate to severe groups and from moderate convalescent to severe convalescent groups. This means that in the progression of disease, the relative composition of ligand genes in myeloid cells and B cells start to differ. On the contrary, when comparing myeloid cells and T cells or B cells and T cells as senders, we found an increase from control to moderate to severe groups and from moderate convalescent to severe convalescent groups. This suggests that as disease progresses, myeloid cells and T cells and B cells and T cells start to recruit similar receivers. These trends also indicates that while the functions of myeloid and B cells as senders alters across COVID-19 progression, their function approaches that of sender T cells.

We also observed a significant decrease in of mean weighted cosine similarity of myeloid cells and T cells as receivers in severe patient group, compared to control and moderate groups. Thus, as disease progresses, the relative composition of receptor genes changes dramatically in myeloid cells and T cells. Overall, we found that myeloid cells and B cells as senders and B cells and T cells as receivers are still the most similar pairs of senders and receivers under this measure, as they have the highest mean weighted cosine similarity in all the five groups.

2.2.7 Diversity of CCC networks in the MIF pathway

To understand the diversity of CCC graphs in different perspectives, we calculated four metrics for measuring graph distances:

(1) Root Euclidean distance between information flow proximity matrices;

(2) apply root Euclidean distance to sender and receiver node-to-node Canberra proximity matrices, and take $d = \sqrt{d^2_{sender} + d^2_{receiver}}$ as the combined distance.

(3) apply root Euclidean distance to sender and receiver weighted cosine proximity matrices, and use similar .

(4) Canberra distance between CCC adjacency matrices.

(1),(2),(3) measures the root Euclidean distances between node-to-node affinity matrices, thus measures the overall node-to-node differences.

Then we calculated variations of network measures within each health condition group (Figure 2.5). In the case of information exchange, we can interpret a higher variation within a condition as a larger difference in network level node-to-node importance. We observed a consistent increase in network diversity from healthy controls to severe patients. These findings indicate that CCC patterns in COVID-19 patients are more diverse compared with healthy controls. Under the aforementioned measures, we also observed a higher variation in the convalescent group from severe symptoms than that in the moderate convalescent group.

2.3 Discussion

As more large-cohort scRNA-seq studies, both in terms of the number of cells and the number of tissue samples, become available, it is important to account for and analyze the individual-sample heterogeneity, both within and across the relevant biological groups, such as disease status. To this end, we derived different diversity measures of CCC and used

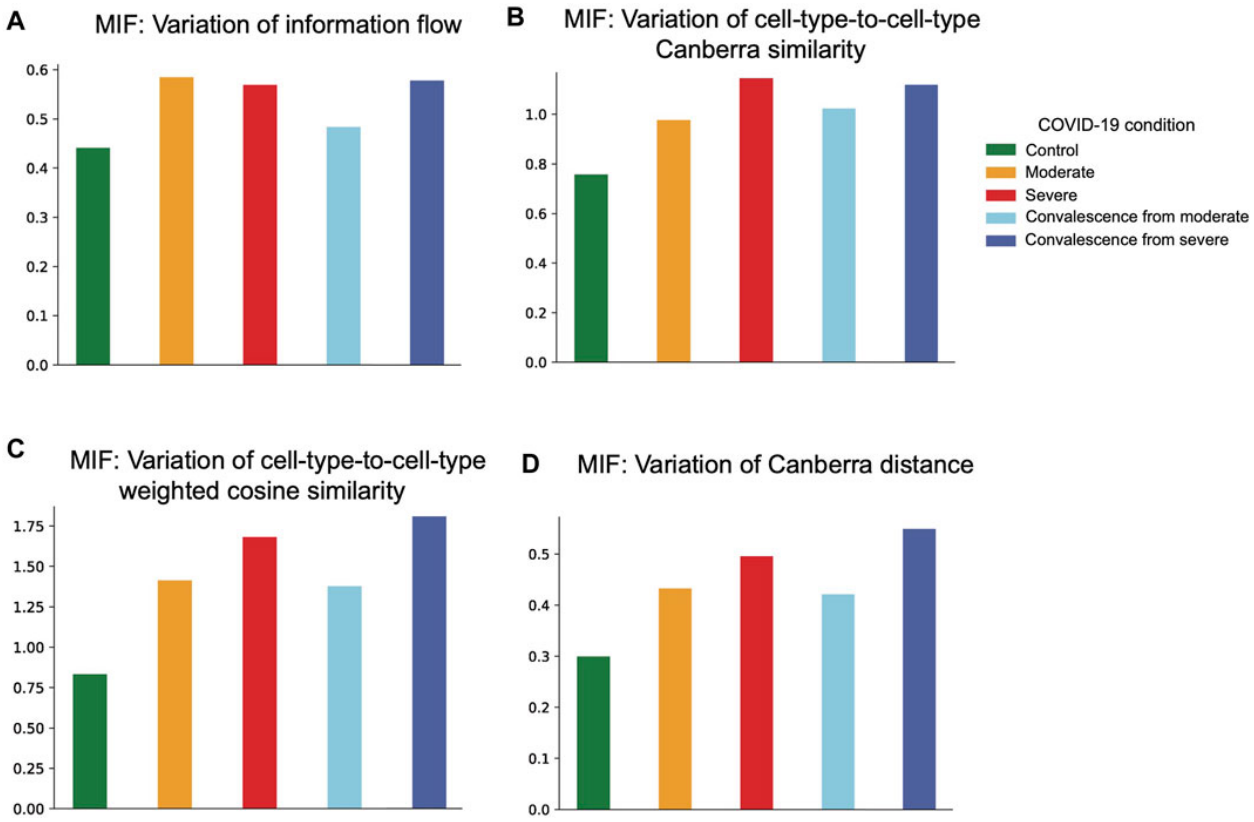


Figure 2.5: **Network variability based on four distance measures.** (a) Network variability based on information flow distance. The variability in moderate and severe groups is higher than that in the control group. The variability in the severe convalescence group is higher than that in the moderate convalescence group. (b) Network variability based on cell-type-to-cell-type modified Canberra distance. There is an increasing pattern of variability from control to moderate to severe groups and from moderate convalescence to the severe convalescence group. (c) Network variability based on cell-type-to-cell-type weighted cosine similarity. There is an increasing pattern of variability from control to moderate to severe groups and from moderate convalescence to the severe convalescence group. (d) Network variability based on graph-to-graph modified Canberra distance. There is an increasing pattern of variability from control to moderate to severe groups and from moderate convalescence to the severe convalescence group.

them to study the diversity of CCC in a scRNA-seq dataset of COVID-19 that was sampled from 181 patients across five conditions, focusing in particular on the MIF pathway. By representing CCC networks as networks, where cell types are nodes and interaction scores are directed, weighted edges, we studied three categories of network features: single node (cell type), node-to-node level, and network-level. We then studied variations across networks in different groups of conditions.

At the node level, we analyzed both sender and receiver variation and entropy. The sender (receiver) variation measures the number of receivers with which (sender) a sender (receiver) cell type interacts, while sender (receiver) entropy measures the uniformity of outgoing (incoming) interaction strengths with different receivers (senders). Together with sender (receiver) degrees, these three diversity measures provide a comprehensive picture of diversity at the cell type (node) level.

Next, we studied pairwise node-to-node or cell-type-to-cell-type level features, which measures either amount of total information exchange from one type to another or the similarity of two senders (receivers). The first measure, information flow from cell type i to j, measures the total amount of information sending from cell type i that is received by j. The second measure, modified Canberra node-to-node similarity, measures the relative difference of interactions between two senders (receivers). The third measure, weighted cosine node-to-node, measures if the senders (receivers) have similar relative composition of ligands (receptors). The modified Canberra and weighted cosine node-to-node similarities can serve as complementary measures of node-to-node similarities. The former provides similarity in relative interaction strengths, and the latter provides similarity in relative composition of relevant ligands or receptors.

Finally, using different metrics between CCC graphs, we studied total variation across graphs in different patient groups comprising conditions. We found that there was a consistent pattern in an increase of variation in COVID-19 groups than in control groups, suggesting a more diverse CCC reaction pattern in diseased patients. The aforementioned analysis of

CCC networks from three angles, namely, single node, node-to-node, and network levels, can be applied to analyze any other pathways in COVID-19. These proposed methods on analyzing CCC diversities within different groups of conditions could also be applied to sub cell types to reveal biologically meaningful results, with the caveat that introducing more cell types can increase the amount of noise from CCC output (Figure A.4-A.8). We noted that using different CCC tools, such as CellCall [112], may result in different CCC output for individual patients, and the overall trend of patient-to-patient variability in CCC will remain, emphasizing the importance of methodologies and analyses that measure the diversity of CCC across cohorts of networks.

One technical limitation of the diversity measures proposed for CCC is that the majority of networks identified signaling pathways were sparse, indicating very few interactions. As cell types were more isolated for these pathways, the node-to-node analyses fail to recognize meaningful patterns. For example, the cosine similarity between two sender cell types is almost exclusively 0 and 1, and the majority of interactions do not overlap. This phenomenon pertains even with weighted cosine similarity for very sparse networks. As a result, the diversity of sparse networks based on these measures cannot be interpreted meaningfully, in the way that dense networks can be. One can use hand-selected features in sparse networks to condense the interaction information, but this may result in loss of biological interpretation after feature selection. In this direction, future work will be needed to address this limitation of sparsity when studying sample-to-sample graph diversity, while retaining biological interpretability of these methods.

2.4 Method Overview

2.4.1 Data Preparation

The scRNA-seq data of COVID-19 progression were downloaded from the NCBI GEO database, accession number GSE158055. The original scRNA-seq data consist of samples

taken from PMBCs, bronchoalveolar lavage fluid (BALF), and sputum. To ensure consistency in biological sample type as well as cell type, we only retained PBMCs. The considered data consist of 181 COVID-19 patients and controls, including 25 healthy controls, 18 patients with moderate symptoms, 43 patients with severe symptoms, 57 patients in the convalescent stage from moderate symptoms, and 38 patients in the convalescent stage with severe symptoms. All single-cell preprocessing and analysis were performed using Seurat [47]. Quality control was performed according to the original study [77]. Specifically, cells with fewer than 1,000 unique molecular identifier (UMI) counts and 500 detected genes were removed, as well as cells with more than 10% of gene counts arising from mitochondrial genes. To remove potential doublets, we also filtered out cells that contained more than 25,000 UMI counts and 5,000 detected genes. In addition, we used Scrublet [107] to identify doublets. The expected doublet rate was set to 0.08, and cells predicted to be doublets were removed. Prior to dimensionality reduction, we determined the 1,500 most highly variable genes. Gene expression counts were normalized to 10,000 counts per cell, log-transformed, and then scaled. We used the original clustering metadata to label B cells, T cells, and myeloid cells.

2.4.2 CCC Analysis by CellChat

To identify the possible cell-cell communications in patients, we employed an R package CellChat [54]. The main function of CellChat is to use scRNA-seq data to compute the strength of cell-cell communication between two group of cells. After clustering all cells in a sample into N different cell-types, CellChat can produce a cell-cell communication network where each node $i \in \{1, 2, \cdots, N\}$ represents a cell-type and directed weighted edges A_{ij} represents the strength of interaction from cell-type i (as sender) to cell-type j (as receiver). In the following we use A to denote the adjacency matrix of this network.

CellChat also provides some built-in functions to quantify and compare relevant biological features of the cell-cell communication networks. The following major functions are used in

our study: information flow of a network, functional and structural similarities between a pair of networks. The information flow of a network is compute by the following formula:

$$I_f = -\big[\log\big(\sum_{i,j=1}^{N} A_{ij}\big)\big]^{-1} \tag{2.1}$$

where log is taken to visualize small differences when a network is nearly full, i.e. the total interaction weights are close to 1. I_f quantifies the total strength of interactions in a network. The functional similarity between two networks quantifies the matching similarity of major sender and receivers of each network, and the structural similarity is used to measure a topological similarity between the structure of two networks without identifying cell-types. To evaluate biological differences across 5 groups using these built-in measures, we first combined all patients in each group as a single sample. Each sample well represents the overall interaction within a group of certain health condition, and the resulting network can be taken as the cell-cell interaction of an average patient in each group.

2.4.3 Measures of Graph Features

In this section, we devise and implement measures of single node and between-node features of a CCC network to compare the diversity of these features across all COVID-19 conditions. Rather than using the methods described in Section 2.2 for aggregated scRNA-seq data, we used CellChat to obtain the CCC network for each individual patient, and the following analysis was performed based on these 181 data points.

Features of Single Node in CCC Network

For each sender or receiver cell type, we consider three measures of diversity:

Node Degree The degree of a sender cell type is the total amount of interactions of the sender cell type.

Node Diversity The diversity of a sender cell type is the number of different receiver cell types it interacts with.

Node Entropy The entropy of a sender cell type measures how balanced or uniform the interaction is[87]. If the interaction of one cell type is uniform, which means if it is evenly spread out with all receiver cell types, then the entropy will be low. On the contrary, if the interaction of cell type A is highly concentrated on one receiver, then the entropy will be high. This measure of how balanced the cell interaction is depends on the relative interaction strengths of one sender cell type between different receiver cell types. Since this measure depends on the proportion of interactions, we first normalize the interactions of each sender cell type i:

$$\tilde{A}_{ij} = \frac{A_{ij}}{\sum_{j=1}^{n} A_{ij}} \tag{2.2}$$

For the edge case of when a sender cell does not interact with any receivers, then it is still 0 after normalization. The entropy of a sender cell type i is then calculated using Shannon entropy:

$$H_{sender}(i) = -\sum_{i=1}^{n} \tilde{A}_{ij} \log(\tilde{A}_{ij}) \tag{2.3}$$

The diversity of receiver cell types are calculated similarly. This entropy is always positive. For edge cases where a sender or receiver have no interactions with the rest of the cell types, we treat its entropy as 0, the lowest possible entropy.

Proximity Measures between Pair of Nodes in CCC Network

A Random Walk Inspired Information Flow Proximity Matrix We can consider the cell-cell communication network as a cell-cell information exchange network. The cell-cell communication probability from cell type i to cell type j, as calculated from CellChat,

can be modeled as the percentage of pathway-relevant information propagating from node (cell type) i to j. Such a propagation can take n steps with $n-1$ intermediate nodes, and in general, $(A^n)_{ij}$ measures the information flow from node i to node j in n steps. Under this framework, we can model the total amount of information exchange matrix S by:

$$S = \sum_{n=0}^{\infty} A^n - I = (I - A)^{-1} - I \tag{2.4}$$

This matrix is the information flow proximity matrix. measures the total information flow from node i to j. This framework is similar to a random walk on a graph. If is a probability transition matrix, then S_{ij} gives the probability of a simple random walk starting with node i will hit node j within finitely many steps.

Canberra Proximity Matrix and Canberra Distance The Canberra distance between two sender cell types and is the Canberra distance between the 2 corresponding interaction vectors $\mathbf{u} = (A_{i1}, \cdots, A_{in}), \mathbf{v} = (A_{j1}, \cdots, A_{jn})$:

$$d_{Can}(\mathbf{u}, \mathbf{v}) = \frac{1}{m} \sum_{i} \frac{|\mathbf{u}_i - \mathbf{v}_i|}{|\mathbf{u}_i| + |\mathbf{v}_i|} \tag{2.5}$$

where the summation is over i such that $\mathbf{u}_i \neq 0$ or $\mathbf{v}_i \neq 0$, and m is the number of summation terms. In the case when $\mathbf{u} = \mathbf{v} = \mathbf{0}$, we define the Canberra distance to be 0.

Canberra distance examines the sum of series of a fraction differences between coordinates of a pair of interaction vectors. This distance is sensitive to small changes when both coordinates are near 0. Low degree of Canberra distance means two interaction vectors are relatively similar with respect to each of their interaction strengths.

The Canberra similarity between $\mathbf{u}, \mathbf{v}$ is defined as $1 - d_{Can}(\mathbf{u}, \mathbf{v})$. The Canberra sender proximity matrix $D_{Can}(s)$ consists of all node-to-node (sender cell type to sender cell type) Canberra similarity. The entry $(D_{Can}(s))_{ij}$ represents the Canberra similarity between sender cell type i and j. The Canberra receiver proximity matrix $D_{Can}(r)$ is defined similarly.

The Canberra distance between 2 cell-cell interaction networks with adjacency matrices A and B are:

$$d_{Can}(A,B) = \frac{1}{m}\sum_{i,j}\frac{|A_{ij}-B_{ij}|}{|A_{ij}+|B_{ij}|} \tag{2.6}$$

the summation is over i such that $A_{ij} \neq 0$ or $B_{ij} \neq 0$, and m is the number of summation terms. In the case when both entries are 0, Similarly, low degree of Canberra distance means two cell-cell interaction networks are relatively similar with respect to each of their interaction strengths.

Weighted Cosine Proximity Matrix Each sender cell type corresponds to a vector of cell-cell interactions from this cell type to all receivers. We are interested in whether the interaction vectors of two different sender cell types are in the same direction. To be more specific, we would like to study if they both interact relatively larger with receiver j than with receiver k. We use a weighted version of cosine similarity. Given two interaction vectors $\mathbf{u} = (A_{i1}, \cdots, A_{in}), \mathbf{v} = (A_{j1}, \cdots, A_{jn})$, the weighted cosine similarity between them is:

$$Cos_{weighted}(\mathbf{u},\mathbf{v}) = \frac{m}{n}\cdot\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \tag{2.7}$$

Where $m = \#\{i : \mathbf{u}_i \neq 0 \text{ or } \mathbf{v}_i \neq 0\}$. In the case when both interactions are 0, we say the cosine similarity between them is 0. This similarity is between 0 and 1. We treat the simultaneous zero entries as missing information, because they do not provide any justification of the similarity. Compared with traditional cosine similarity, the advantages of weighted cosine similarity are: m/n is the measure of credibility of similarity, the amount of information that the interactions reveal. In the case when interaction vectors $\mathbf{u}, \mathbf{v}$ are generally sparse, less informative interaction vectors display less similarity.

The weighted cosine sender proximity matrix consists of all node to node (sender cell type to sender cell type) weighted cosine similarity. The ij^{th} entry represents the weighted cosine

similarity between sender cell type i and j. The weighted cosine receiver proximity matrix is defined similarly.

Diversity Measures of Group of CCC Networks

We use the root Euclidean distance to measure the distance between node-to-node proximity matrices $S^{(1)}, S^{(2)}$:

$$d_r(S^{(1)}, S^{(2)}) = \sqrt{\sum_{i,j=1}^{N} \left(\sqrt{S_{ij}^{(1)}} - \sqrt{S_{ij}^{(2)}}\right)^2} \tag{2.8}$$

Compared to the normal Euclidean distance, d_r is more sensitive to entries S_{ij} that are close to 0, rather than giving equal weights for all entries. This property allows root Euclidean distance to display considerable differences when the network connection is relatively sparse. Based on the root Euclidean distance, we study the variation of cell-cell interaction graphs under different disease conditions. For a group of patients with CCC networks $A^{(1)}, A^{(2)}, \cdots, A^{(n)}$, we use the following definition as the measure of variance within the group:

$$Var(A^{(1)}, A^{(2)}, \cdots, A^{(n)}) = \min_{i=1,\cdots,n} \sum_{j \neq i} d_r(S^{(i)}, S^{(j)}) \tag{2.9}$$

Where $S^{(i)}$ is the (Information Flow, Canberra or Weighted Cosine) proximity matrix of $A^{(i)}$.

Chapter 3

Detect Hierarchical Structures in Cell-cell Communications via Single-cell and Spatial Transcriptomics Data

3.1 Introduction

In a multicellular organism, biological functions are performed by cells coordinating via cell-cell communication (CCC). Cells communicate with each other by secreting molecular signals, which are received by nearby cells. CCC instructs crucial cellular functions, including cell development, homeostasis, and immune responses in disease [4, 105, 1, 3]. CCC occurs at different biological scales. While some communications occur over a wide range of cells, others may occur only among a small subset of cells. The first example of such a hierarchy in CCC is related to cell types at varying scales. Specifically, there may be sub-cell types within a cell type. T follicular helper cells, which are a subset of CD4+ T cells, have been found to secrete IL-4 and IL-21 ligands, which are received by B cells to promote cell prolif-

eration and B cell differentiation into plasma B cells or germinal center B cells[30]. Another ligand-receptor interaction, CD40L-CD40 interaction occurs at a wider range of cells, from CD4+ T cells to B cells. This interaction is crucial for T-cell-dependent B cell proliferation and differentiation[56]. These two examples of ligand-receptor interactions demonstrate the possibility of a hierarchy among different CCC activities. The second example of a possible hierarchy in CCC is determined completely by the cells where ligands/receptors involved in the interaction are expressed. For example, there are genes expressed during a specific phase of a cell cycle among many different cell types, and they can perform important biological functions such as regulating cell growth, DNA replication, and cell division[39]. As a specific example, senescent cells can induce paracrine effects on other within their tissue microenvironment via the senescence-associated secretory phenotype, which is characterized by the secretion of a wide array of inflammatory cytokines, chemokines, and growth factors[100]. Therefore, considering cells that express a certain ligand/receptor, while relaxing the restriction that they should belong to a certain cell type, is biologically more relevant when studying CCC.

Therefore, identifying which groups of cells exhibit hierarchical CCC structure could increase one's understanding of the structured biological functions that occur among cell groups due to inherent structures within ligand-receptor interactions. Such an understanding of structures within CCC will shed light on the systematic mechanism of cellular events, such as cell development and differentiation, as well as orchestration of different functions across various biological conditions such as inflammation, wound healing, and cancer.

Recent advances in single-cell RNA-sequencing and spatial transcriptomics technologies have paved the way to understanding CCC, and a rich number of methods have been developed to infer CCC based on scRNA-seq or spatial transcriptomics data[54, 35, 106, 72, 16, 14, 17, 50, 49, 76, 34, 18, 110, 60]. Current tools for modeling CCC typically operate under specific assumptions and focus on interactions among groups of cells that are predefined. However, this approach has two significant drawbacks. Firstly, it overlooks the intricate cell

structures shaped by ligands and receptors, crucial components in CCC. More importantly, it only accounts for CCC between cell groups at a one level of clustering, which overlook the varying scales of cell states, i.e. there may be substates within clusters, and further substates within those substates. In the example of the interaction between T follicular helper cells and B cells through secretion of IL-4 and IL-12 ligands, if we only consider one scale of clustering, where T cells are a cluster, the two interactions sent from T follicular helper cells, and CD4+ T cells respectively, will both be assigned as interactions from T cells. A coarse scale of clustering may not be able to ascribe the interactions to the correct cell types as precisely as desired. However, focusing solely on the finest scale of clusters may cause us to miss the broader "big picture" functions of T cells and B cells. Therefore, it is essential to consider the cellular hierarchy rather than just a single level of clusters.

The key to overcoming these limitations is to develop a method that can analyze CCC between cell groups at multiple clustering resolutions defined by expression of signal ligands and receptors. Hierarchies in cell groups occur throughout biology. For example, a hierarchy naturally develops during cell differentiation, when a subcluster of cells start to perform more specialized functions among a larger group of cells. Consequently, we can define hierarchical structures within CCC with respect to two complementary perspectives. First, from a global perspective, the hierarchical structure divides all cells into clusters at different scales, where a cell group at a coarser scale indicates a general cell group, and a cell group at a finer scale indicates a more differentiated subcluster. A ligand-receptor interaction can be specific to a pair of cell groups belonging to any scale along this hierarchy, and thus the global hierarchy of CCC can be induced from this hierarchy of cells. Second, from a local perspective, cells expressing a ligand A may be a subset of cells expressing ligand B (with a similar analogy for signal receptors). If we analyze ligands and receptxors with respect to their expressing cell groups, some ligands or receptors will form natural tree structures based on the overlap or lack thereof between the cell groups that express these ligands. Several recent studies investigated the existing hierarchical patterns in scRNA-seq datasets[67, 48, 57, 89, 38].

However, they focus on the hierarchy from a global perspective, and do not study the local hierarchy with respect to CCC.

Thus, we propose CrossChat, which detects and analyzes both global and local hierarchical structures within CCC generated by ligand-receptor interactions using single-cell and spatial transcriptomics. CrossChat consists of two core methods, CrossChatH and CrossChatT, and their respective extensions to spatial transcriptomics data, CrossChatH-S and CrossChatT-S. CrossChatH detects a global hierarchical communication structure based on a hierarchical community detection method. CrossChatT is able to detect multiple local hierarchical structures based with respect to ligand gene expression or receptor gene expression using a tree detection method. CrossChat presents several major advantages. First, it detects structures within CCC that do not rely on predefined cell type annotations. Second, it detects the structure of CCC with respect to ligand and receptor gene expression, and thus more accurately reflects the inherent structures within CCC. Third, it provides a comprehensive view on hierarchical structures within CCC with respect to both global and local perspectives. To the best of our knowledge, CrossChat is the first method to detect and analyze hierarchical structures within CCC. CrossChat is available as an open-source Python package, that provides inference, visualization, and downstream applications of hierarchical structures within cell-cell communications.

3.2 Methods

3.2.1 Overview of CrossChat

From a global perspective, hierarchical structures CCC can be uncovered by hierarchically clustering cells according to their gene expression similarity or their similarity in ligand or receptor gene expression. Some ligand-receptor interactions may occur between smaller cell groups in which cells are highly similar, while other interactions may occur between larger cell groups where cells are less similar (Figure 3.1a). From a local perspective, hierarchical

structures in CCC can be uncovered by analyzing, for example, whether a group of cells that secrete a signal ligand contains a smaller subset of cells that secrete a secondary signal ligand. Similar structures may exist with respect to signal receptors or ligand-receptor pairs (Figure 3.1b). Given these hierarchical structures within ligands or receptors, we can build hierarchical relations between ligand-receptor interactions (Figure 3.1c). These hierarchical structures of ligands, receptors, or interactions can be visualized using trees (Figure 3.1d-f). CrossChatH investigates the global hierarchical structure within CCC based on hierarchical clustering of cells. It first detects hierarchical cell groups based on gene expression or ligands/receptors expression similarity, and then calculates CCC activity between clusters (Figure 3.2a). The input to the method either a log-normalized gene expression matrix or a raw count matrix. For a raw count matrix input, it will first log-normalize it. It then calculates a cell-cell similarity graph based on cosine similarity, which is then used to produce a K-nearest neighbor graph, where the nodes represent the cells, and the edge weights represent the similarity of cells. Using PyGenStability[5], a multi-resolution community detection method with generalized Markov Stability[84], we obtain a hierarchy structure within cells. Next, we obtain an ordered list of specific ligand-receptor pairs, ranked by their specificity to ligand and receptor clusters in the hierarchy. For each ligand-receptor pair, CCC is calculated using CellChat9. We tested the robustness of hierarchical clustering of CrossChatH to noise, sparsity, and normalization, selection of K in K-nearest neighbor graph construction (Figure B.1a-d).

CrossChatT investigates the local hierarchical structure with respect to ligands, receptors, or ligand-receptor interactions using Bron-Kerbosch, a graph search algorithm[13] (Figure 3.2a, Figure B.2). The input to the method is a raw gene expression matrix, which is then subsetted to consider only ligand and receptor genes and then binarized to construct a support matrix. For the set of ligands (receptors), we connect any pair of ligands (receptors) based on whether they adhere to the principle of being disjoint or inclusive. This procedure generates a gene-gene relationship graph where graph nodes represent genes, and edges

represent gene-gene relationships. The Bron-Kerbosch algorithm is used on this gene relationship graph to find maximal complete subgraphs (largest subgraph in which each pair of nodes is connected). Each subgraph represents a local hierarchical structure within the set of ligands/receptors, where each pair of ligands/receptors are either disjoint or inclusive, forming a tree structure. For each pair of ligand tree and receptor tree which has at least one known ligand-receptor interaction, we calculate CCC using CellChat. We tested how different binarization thresholds could affect the detection of local hierarchies (Figure B.1e-h). We found that the number of detected ligand trees decrease as we increase the binarization threshold (Figure B.1e). Howver, we observed that neither the number of detected receptor trees (Figure B.1f) nor the average number of ligands/receptors in ligands/receptors trees (Figure B.1g–h) change significantly as the binarization threshold changes. CrossChat provides functionality for downstream analysis of these hierarchical structures extracted from CCC (Figure 3.2c). First, CrossChat groups signaling pathways or ligand-receptor pairs based on their hierarchical distribution of ligands/receptors by calculating the cosine similarity between ligands/receptors distribution over hierarchical clusters, which is weighted by Markov times (see Methods). Second, CrossChat provides a global view of interactions between ligands trees and receptor trees by visualizing interactions between a combination of ligands/receptors trees. Third, CrossChat can measure the importance of a ligand or receptor in the hierarchy of ligands/receptors by ordering them by their frequency of occurrence within the sets of ligands/receptors trees.

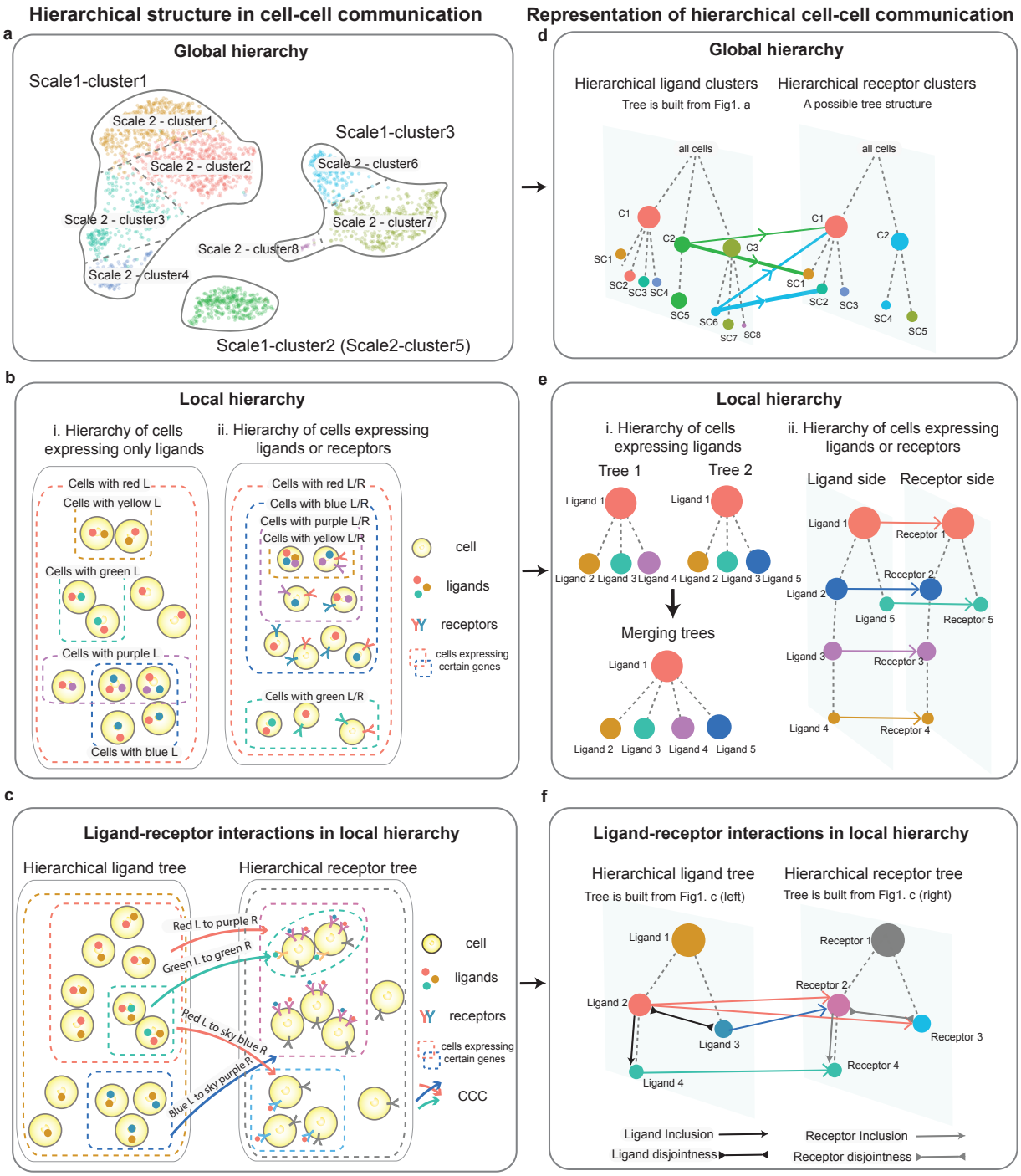


Figure 3.1: **Hierarchical structures in cell-cell communications.** (a) A hierarchical clustering of cells. Cell groups at scale 2 are subclusters of cell groups at scale 1.

Figure 3.1 (continued). (b) Hierarchical structures within ligands, receptors, or union of ligands and receptors. (i) In the hierarchy of ligands, all cells express red ligands, some cells express yellow ligands, and some express green ligands, etc. (ii) In the hierarchy of ligand-receptor unions, all cells express either the red ligand or receptor. Within the group of cells expressing the yellow ligand or receptor, there are also cells expressing either the purple ligand or receptor, and blue ligand or receptor.
(c) Hierarchical structures of ligand-receptor interactions. There are four hierarchical relationships of ligand-receptor interactions: ligand inclusion, ligand disjointness, receptor inclusion, and receptor disjointness.
(d) Representations of cell-cell communication between hierarchical clustering of cells based on ligands (left) and receptors (right). The node sizes correspond to the number of cells in the groups, and the edge widths represent interaction strengths.
(e) Tree representations of hierarchical structures within ligands, receptors, or ligand-receptor unions.
(f) Representations of hierarchical relations between ligand-receptor interactions.

3.2.2 CrossChatH: hierarchical clustering

CrossChatH investigates the global hierarchical structure of cell–cell communication (CCC) by performing hierarchical clustering at the single-cell level. It first identifies hierarchical cell groupings based on similarities in gene expression or ligand–receptor profiles, and then quantifies CCC activity between the resulting clusters.

0. Preliminaries

Single-cell data are commonly represented as a gene expression count matrix, capturing the expression levels of over 20,000 genes across individual cells. To prepare the data for downstream analysis, initial preprocessing steps are applied to remove low-quality or uninformative genes. This is followed by log-normalization and scaling of the expression values. The resulting matrix is a normalized gene expression matrix $M \in R^{n \times m}$, where n is the number of cells and m is the number of genes. Principal Component Analysis (PCA) is then performed on the scaled data to reduce dimensionality and capture the most informative variation.

1. KNN graph construction

Using latent embeddings generated by PCA, we first use cosine similarity to construct a graph $G = (V, E)$. The nodes V of the graph are cells, and the edges $E \subseteq V \times V$ represent the similarity between the cells. The graph G is an undirected and weighted graph. Denote the adjacency matrix of the graph G as A. Denote the diagonal matrix D as the degree matrix of A, where $D_{ii} = \text{degree}\,(v_i)$, where v_i denotes the i^{th} node of the graph.

2. Random walks on graphs

Random walks on graphs are stochastic processes describing the trajectory of a particle or entity that moves randomly from node to node along edges within a graph. These walks provide a powerful framework for modeling numerous phenomena, including diffusion processes on graphs.

Definition 1. Given a graph $G = (V, E)$ with adjacency matrix A and degree matrix D, we define the following Laplacian matrices:

- The *combinatorial Laplacian* L:

$$L = D - A. \tag{3.1}$$

- The *symmetrically normalized Laplacian* $\mathcal{L}$:

$$\mathcal{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2}. \tag{3.2}$$

- The *random walk Laplacian* L_{rw}:

$$L_{\text{rw}} = D^{-1} L = I - D^{-1} A. \tag{3.3}$$

Definition 2. Given a graph $G = (V, E)$ with adjacency matrix A, we define the continuous

time transition matrix as:

$$P(t) = e^{-tL_{rw}} \tag{3.4}$$

The random walk transition matrix, $D^{-1}A$, has stationary distribution π, where π_u is the limit probability of a random walk that ends up in vertex u as time t tends to infinity.

3. Community Detection and Markov stability

Community detection refers to the task of identifying groups or clusters (also known as communities) within a graph, where nodes inside each community exhibit higher connectivity or similarity among themselves than with nodes outside their community.

Random walks are widely utilized for community detection tasks due to their intuitive interpretation and computational efficiency. The core idea is that a random walker tends to remain longer within highly connected subgraphs (communities) compared to less densely connected areas. By analyzing the patterns or probabilities associated with random walks, communities can be naturally uncovered.

Markov Stability extends the random-walk framework by explicitly considering how community structure emerges across different time scales of diffusion processes. Specifically, it quantifies the tendency of a random walker to remain within the same community after a given Markov time, enabling detection of robust communities at various resolutions—from detailed local groups at short time scales to broader modular structures at longer time scales. This multiscale approach makes Markov Stability especially powerful for uncovering hierarchical or nested community structures that standard single-scale methods may overlook.

Definition 3. Given communities $C_1, \ldots, C_k$ present in the graph G and an associated random walk transition matrix $D^{-1}A$ with stationary distribution π, the Markov stability[31]

(MS) of community C_i is defined as

$$\mathrm{MS}(C_i, t) = \sum_{u,v \in V} (P(t)_{uv} - \pi_v)\, \delta(c_u, c_v) \tag{3.5}$$

where the sum is defined over all vertices V. Here, δ is the Kronecker delta defined such that $\delta(c_u, c_v) = 1$ if vertices u and v are in the same community and $\delta(c_u, c_v) = 0$ otherwise. $P(t)$ is the continuous time transition matrix (3.4).

Two partitions are considered equivalent if they differ only by a permutation of community labels. Intuitively, for a given way of partitioning the graph into communities, the Markov stability measures the overall likelihood that the random walker stays inside its community at time t. If we fix the communities, this likelihood will decrease as t grows, since the random walker tends to explore more other communities rather than staying in its own. For each t-value, we use the Louvain algorithm to find communities that maximize the MS at this t-value. Due to the previous fact that the MS decreases with t for fixed communities, we tend to have fewer (and larger) communities when maximizing the MS for larger t-values, converging to the case where one community contains the whole graph.

4. Variation of information

The robustness of partitions of the graph G found at each time scale t is assessed using *variation of information* (VI), which measures the distance between two partitions.

Definition 4. Given two partitions of the graph G, $\{X_i\}_{i=1}^{k}$, $\{Y_j\}_{j=1}^{l}$, given

$$p_i = \frac{|X_i|}{n}, \quad q_j = \frac{|Y_j|}{n}, \quad r_{ij} = \frac{|X_i \cap Y_j|}{n},$$

where n is the number of vertices in G. Then, the variation of information between the two

partitions is defined as:

$$\mathrm{VI}\left(\{X_i\},\{Y_j\}\right) = -\sum_{i,j} r_{ij}\left[\log\left(\frac{r_{ij}}{p_i}\right) + \log\left(\frac{r_{ij}}{q_j}\right)\right]. \tag{3.6}$$

Before calculating the variation of information at each time t, we first transform t to lie on the $\log_{10}$ scale, setting the range of t to be from 10^{-1} to 10^4. Next, we pick 100 evenly spaced time scales, l_i, ranging from -1 to 4, and select scales $t_i = 10^{l_i}$.

At each time scale t, we run the Louvain algorithm 100 times to generate 100 optimal partitions, denoted as $\{P_i\}_{i=1}^{100}$. Next, we calculate the average pairwise variation of information by

$$\mathrm{VI}_g = \frac{1}{100\times 100}\sum_{i,j}\mathrm{VI}(P_i,P_j).$$

For each time scale t, we select the partition with the highest Markov stability as the representative partition P_t. Then, for each pair of time scales t and t', we calculate the pairwise variation of information $\mathrm{VI}_{P_t,P_{t'}}$, which forms a 100×100 matrix, VI_P.

Intuitively, the VI_P can be viewed as a graph over the lattice $\{(i,j) \mid 1\le i,j\le n\}$. The diagonal elements of VI_P are 0, and the off-diagonal elements are non-negative. Thus, each diagonal entry can be viewed as the center of a basin of the graph. If the representative cluster at time scale t is stable, then its pairwise VI with representative partitions at neighboring time scales will be small, thus forming a large flat basin. The optimal scales are selected by thresholding the width and depth of the basin. We then take the partitions at the selected optimal scales.

5. Hierarchical clustering

Given a graph G with n vertices, the hierarchical clustering algorithm finds a series of optimized graph partitions at different time scales t by maximizing the Markov Stability function (3.5) . The scale parameter t, which is referred to as Markov time, regulates the

coarseness of the partitions. Louvain is performed at each Markov time t to find communities that maximize Markov Stability.

Next, to detect robust multiscale clusterings, we follow the procedure in section 4 (Variation of information) to finalize the detected scales of clustering.

6. Obtain ordered list of specific ligand-receptor pairs

First, at each scale, we use Wilcoxon rank-sum test to determine ligands and receptors that are associated significantly to a cluster, assigning the ligands and receptors to the cluster with the lowest associated p-value. For a ligand (receptor) multi-unit, we calculate the geometric mean of each ligand (receptor) in the complex to measure ligand (receptor) multi-unit expression. Then its p-value is calculated based on the ligand (receptor) multi-unit expression. Next, we only retain the ligands and receptors with associated p-values such that $p < 10^{-3}$. We reason that these filtered ligands and receptors are specific to its assigned clusters, due to their low p-values. Then, we retain ligand-receptor pairs for which both ligand and receptor are inferred marker genes of certain clusters, forming the specific ligand-receptor interaction list. We assign a specificity score each ligand-receptor pair to capture the specificity of the interaction to sender and receiver cluster, calculated by the products of p-values associated with the ligand and receptor, respectively. Finally, we order the specific ligand-receptor pairs their calculated scores.

7. Multiscale similarity of ligand-receptor pairs and pathways

To better illustrate the relationship among multiple ligand-receptor pairs or pathways, we calculate a weighted cosine similarity between each ligand-receptor pair or between pathways of interest. For each ligand, its mean gene expression in each cluster at varying resolutions is calculated and summarized into a vector for each scale:

$$\text{similarity}(L_1, L_2) = \sum_{i=1}^{n} \frac{1}{\log(T_i) + 1} \cdot \text{cosine_sim}(u_i, v_i),$$

where n is the number of scales, T_i is the Markov time of scale i, and u_i and v_i represent the mean gene expression vectors of L_1 and L_2, respectively, across clusters at scale i. We calculate the similarity between receptors in a similar manner. We calculate the similarity between two ligand-receptor pairs as the sum of pairwise ligand similarity and receptor similarity.

For two pathways, we sum the interacting ligand-receptor pairs to obtain a pathway ligand gene expression and a pathway receptor gene expression in each cluster at varying resolutions. The weighted cosine similarity is then calculated in a similar manner by aggregating their ligand similarity and receptor similarity. Then, we apply UMAP to the similarity matrix and use K-means to cluster the ligand-receptor pairs or pathways.

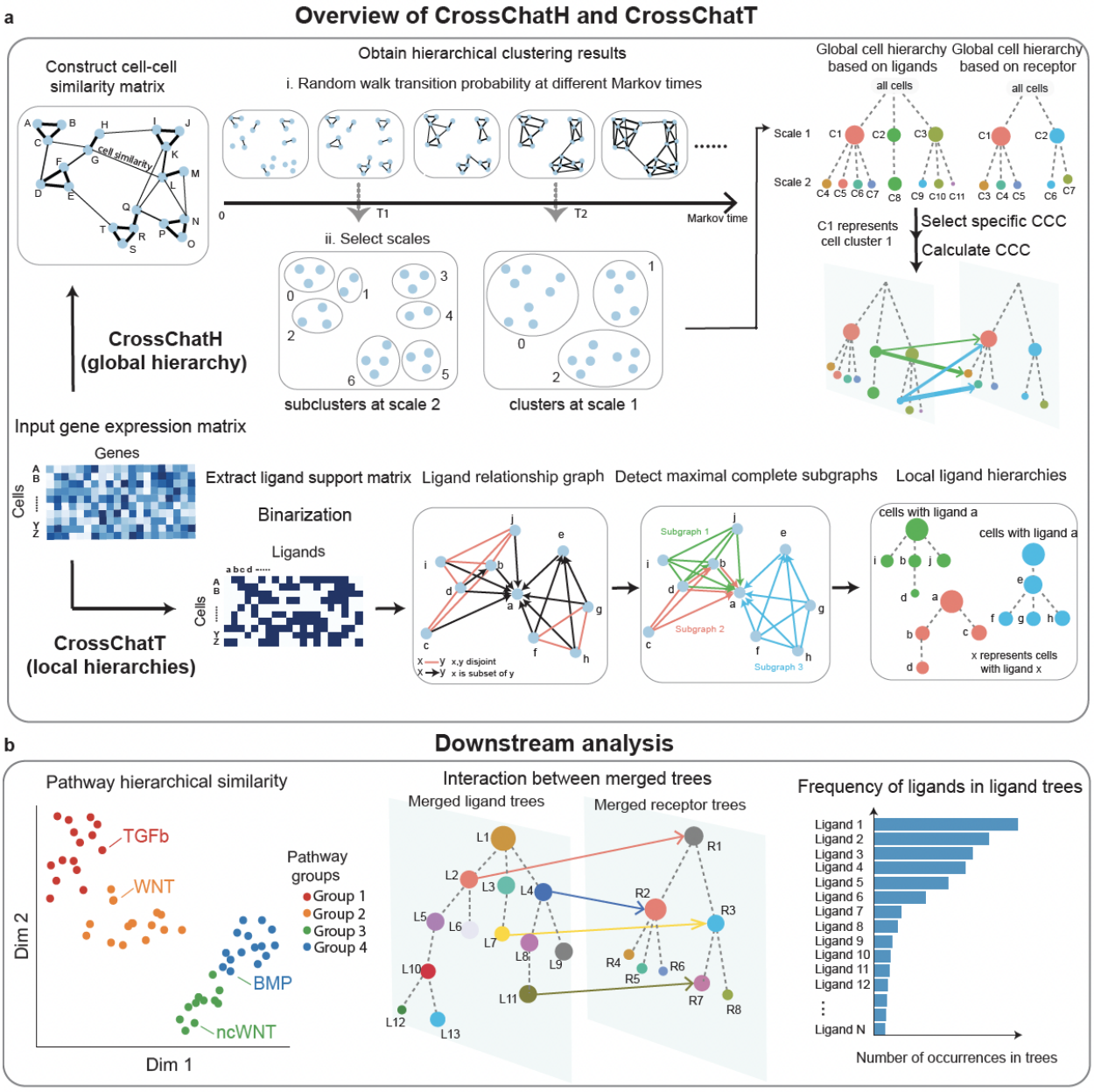


Figure 3.2: **Method overview of CrossChatH and CrossChatT.** (a) Method overview of CrossChatH and CrossChatT.
(i) **CrossChatH: hierarchical clustering.** The input is a gene expression matrix. CrossChatH calculates a cell-cell similarity matrix and uses a random walk-based method on the similarity matrix to obtain hierarchical clustering results. After obtaining ligand clusters and receptor clusters, it runs CellChat to calculate cell-cell communication between clusters.
(ii) **CrossChatT: tree detection of ligand-receptor interactions.** The input is a gene expression matrix, which is then binarized to form a ligands/receptors support matrix. A gene-gene relationship graph is constructed, where gene pairs whose support is either disjoint or inclusive are connected. The Bron-Kerbosch algorithm is used on the gene relationship graph to find maximal complete subgraphs. Each subgraph represents a tree structure of ligands/receptors.

Figure 3.2 (continued). (b) Downstream analysis of CrossChatH and CrossChatT: (i) Similarity of pathways based on ligand/receptor distributions over hierarchical clusters. (ii) Visualization of ligand-receptor interactions between the union of ligand/receptor trees. (iii) Frequency of ligands/receptors in all ligand/receptor trees.

3.2.3 CrossChatT: CCC Tree search

CrossChatT investigates the local hierarchical structure with respect to ligands, receptors, or ligand-receptor interactions.

In order to find ligands/receptors whose supports form a hierarchical tree structure, we construct trees such that, in each tree, all pairs of ligands/receptors are either disjoint or inclusive in terms of their gene support, i.e., the cells that express these ligands/receptors. We impose that each tree has to be maximal in the sense that no other ligand/receptor can be included in this tree. We first binarize the gene expression matrix in order to find the support for each ligand/receptor (Figure B.2). Next, we model that two ligands/receptors are connected if their relationship (with respect to their supports) is either disjoint or inclusive. Two ligands/receptors are disjoint from each other if the cells expressing them are exclusive. One ligand is a subset of another ligand if the cells expressing ligand 1 also expresses ligand 2. Thus, we form a gene relationship graph based on these relationships between ligands/receptors. In a tree of interest, all pairs of ligands/receptors are either disjoint or inclusive, thus are connected. Such a tree structure forms a complete graph, where nodes represent ligands/receptors, and every pair of nodes has an edge connecting them. In order to find largest trees of ligands/receptors in which any other ligand/receptor is either intersecting with one ligand/receptor in the tree or is disjoint with any ligand/receptor in the tree, we use Bron-Kerbosch graph search algorithm to find all maximal complete subgraphs32. Each complete subgraph may be one tree itself or contains multiple disjoint trees. Finally, we check each maximal complete subgraph and extract all individual trees from them (Figure B.2).

3.2.4 CrossChatS: Incorporating spatial information

CrossChatH-S: Spatial position of both directions is first scaled into 0 to 1, and is concatenated to the PCA embeddings of cells. The concatenated PCA embeddings of cells that contain the spatial information of cells are used to calculate the similarity of cells and build the KNN graph on cells. We calculate spatial CCC using COMMOT22.

CrossChatT-S: For any ligand or receptor, we restrict its support such that only ligands or receptors whose nearby spots express its receptors/ligands are kept in its support. Then the filtered ligands/receptors support are used to detect tree structures. Since the resolution and capture ability are different for different spatial sequencing methods, we allow users to modify their own spatial range based on their input data.

3.2.5 Incorporation of CellChat and COMMOT

After CrossChatH or CrossChatT obtains the clusters at multiple resolutions, CellChat (for nonspatial datasets) or COMMOT62 (for spatial datasets) is used to calculate the interactions between the clusters.

CellChat deduces significant CCC activity from scRNA-seq data between identified cell groups. It computes an interaction score using mass action kinetics, which represents the probability of CCC. This process integrates gene expression data with existing knowledge about the interactions among signaling ligands, receptors, and their cofactors.

COMMOT infers CCC according to predefined ligand-receptor pairs by solving a global optimization problem. This method accounts for higher-order interactions among multiple ligand and receptor species. It introduces collective optimal transport to determine optimal transport plans for all species pairs simultaneously, allowing for interdependent couplings that are not possible with traditional optimal transport.

3.2.6 Validation of CrossChat using simulated dataset and COVID-19 dataset

Synthetic data generation

To benchmark the ability of CrossChatH to detect specific hierarchical CCC, we simulated a dataset with 1000 cells and 10,000 genes, among which 1000 are marker genes, and three clustering scales (see Figure 3.3a for details). In each scale, there are two clusters, four clusters, and eight clusters respectively. This simulated dataset is based on our analysis of the characteristics of a real biological dataset of PBMC from 10X Genomics (https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz). The dataset contains 2700 single cells and 13,714 genes.

Step 1:

We normalize the gene expression counts of the PBMC dataset so that each cell has the same number of total counts, which is calculated as the median of the total counts across all cells. Using the cell type annotations provided by Seurat70, we find all marker genes using default parameters by Seurat and order marker by log fold change (based on "one vs. rest" comparison). We observed that, in the normalized PBMC data, the top 1,371 marker genes (10% of all genes) are among the top 20% of all genes in terms of their average expression across cells. Therefore, from 13,714 genes in the PBMC dataset, we randomly choose 1000 genes from the top 20% of genes ranked by average expression. We assign 200, 100, and 50 marker genes, respectively, to each cluster from the coarsest scale to the finest scale (Figure 3.3a). There are 400, 400, and 200 marker genes in total for each scale of clusters.

Step 2:

We set the mean gene expression of our 1000chosen markers as the mean gene expression of the chosen 1000 genes from top 20% genes from the PBMC dataset. Next, at each scale, we set the fold change of marker genes as the median fold change of the top 400, 400, and 200 differentially expressed genes of the randomly chosen 1000 genes, respectively. Using

the mean gene expression and median fold change of each gene, we calculated its mean gene expression both inside and outside of the cluster where the marker should be assigned.

Step 3:

As the Negative Binomial is the most common distribution used to model RNA-seq count data71, we generated simulated gene expression counts for the 1000marker genes from a Negative Binomial distribution with the mean set as the previously calculated mean gene expression counts, and a dispersion parameter of 0.1 (default value chosen by Splatter71). Next, we randomly choose 9000mean gene expression values from the lower 80% of genes ranked by mean gene expression from the PBMC dataset, and simulate gene expression counts for 9000non-marker genes from a Negative Binomial distribution, using the chosen mean gene expression counts and dispersion parameter of 0.1.

Validation of CrossChat

To benchmark the ability of CrossChatH to detect specific CCC, we simulated a dataset with 1000cells, and 10,000genes, among which 1000 are marker genes, and three clustering scales. In each scale, there are two clusters, four clusters, and eight clusters respectively. We reason that, as more marker genes will be inferred for larger clusters, we assign 200, 100, and 50 marker genes, respectively, to each cluster from the coarsest scale to the finest scale (Figure 3.3a). There are 400, 400, and 200 marker genes in total for each scale of clusters. We apply CrossChatH to the log-normalized simulated count matrix to assess its ability to detect the correct marker genes of clusters at different scales. Among the top 1000 selected differentially expressed genes, we found that 789 of them are our assigned marker genes and are accurately assigned to its differential cluster, reaching a success rate of 78.9%. Indeed, at the two finer scales, the success rate reaches 86.5% and 86.0% respectively. Genes that are not ground truth marker genes tend to be assigned to larger clusters than smaller clusters, so it is expected that clusters at coarser scales will have lower success rates. As we increase the number of selected differentially expressed genes from 1000 to 2000, the percentage of

correctly identified differentially expressed genes increases to 91.8%, where the success rate at scales 1, 2, and 3 reaches 96.5%, 95.0%, and 86.2%, respectively (Figure 3.3b). In addition, CrossChat almost perfectly identifies multi-resolution clusters at the simulated scales 1, 2, 3, achieving an Adjusted Rand Index (ARI) of 0.99, 1, and 1, respectively, where an ARI of 0 indicates no overlap while an ARI of 1 indicates complete overlap (Figure 3.3c).

The ability of CrossChatH to detect hierarchical marker genes across scales demonstrates that CrossChatH can find hierarchical structures in cells when using all genes as input. CCC can then be visualized between these hierarchical clusters. When using ligands/receptors for clustering, CrossChatH can detect hierarchical clusters that have statistically significant CCC patterns. The detected markers for each cluster in the hierarchy of cells are ligands/receptors, so it can be interpreted as a hierarchical CCC.

Furthermore, we validated the detected clusters using PBMC (peripheral blood mononuclear cells) data from COVID-19 patients. Two independent hierarchical clustering procedures are performed based on ligands only and receptors only. The boxplot shows the distribution of distance of CCC pattern of a certain cluster with the rest of the clusters. We observe that cell-cell communication patterns are robust to subsampling within each hierarchical ligands/receptors group (Figure 3.1d).

The algorithm used by CrossChatT to find all maximal complete subgraphs is Bron-Kerbosch[13]. Finding hierarchical ligands/receptors trees is equivalent to finding maximal complete subgraphs in the graph of ligands/receptors that we constructed (see "CrossChatT: CCC Tree search" in Methods). To validate CrossChatT, we generated 1000 cells that expressed 100 genes. We distributed the gene expression counts so that there are at least five trees as shown in the figure (Figure 3.1e). There are 15 genes generating these five trees. For each of the other 85 genes, we randomly chose 500 cells to express them. We ran the experiment 1000 times and found that CrossChatT is able to detect all five trees we generated in each trial (Figure 3.1e).

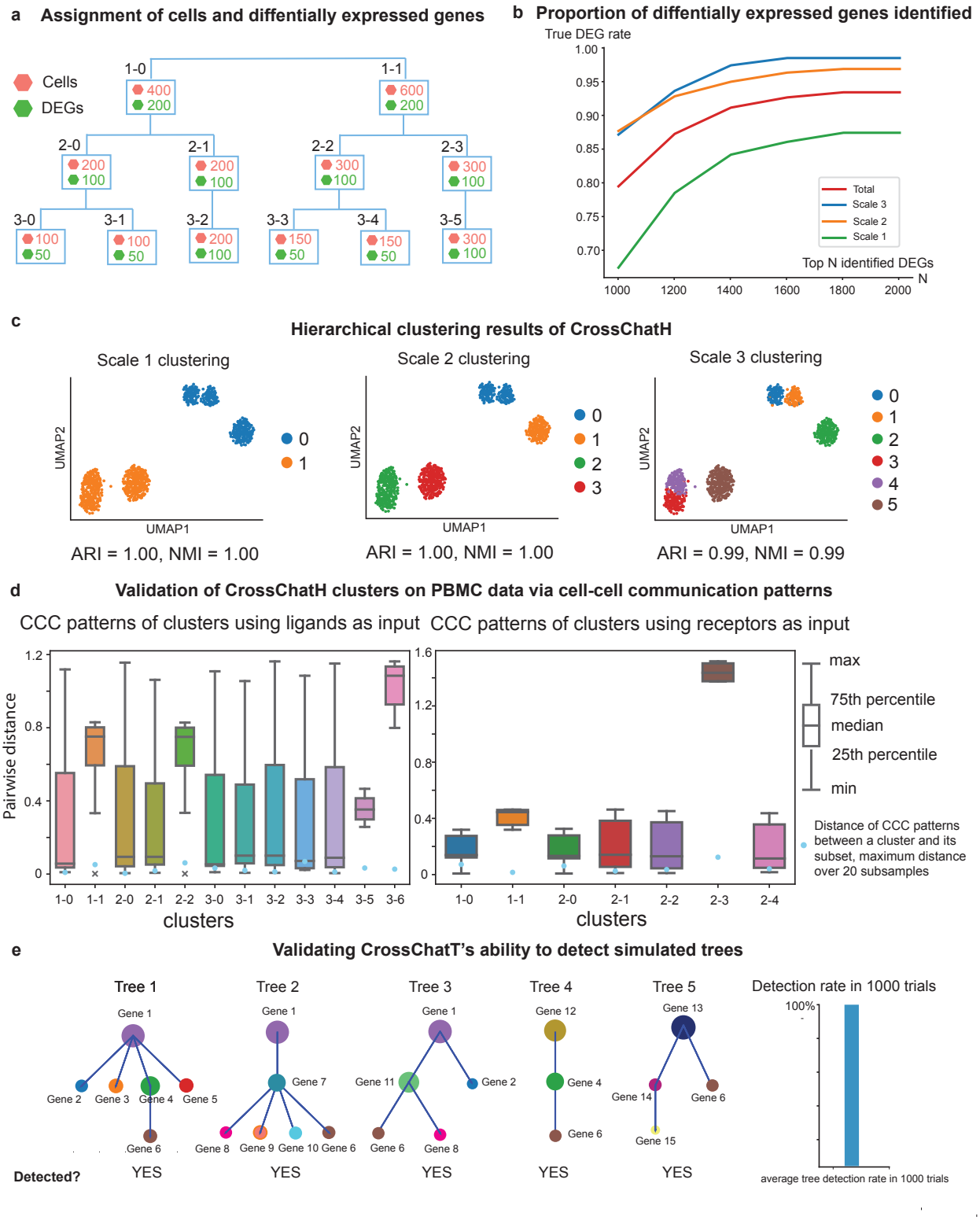


Figure 3.3: **Validation of CrossChatH.**

Figure 3.3 (continued). (a) Synthetic data preparation. Differentially expressed genes (DEGs) are assigned to each cluster. (b) Proportion of DEGs identified at scale 1, scale 2, scale 3, and across all scales.
(c) The hierarchical clustering results of CrossChatH on synthetic data. Clustering at each scale aligns well with the original hierarchical cell group assignments.
(d) Validation of CrossChatH clusters using PBMC data from COVID-19 patients. Two independent hierarchical clustering procedures are performed based on ligands only and receptors only. Boxplot shows the distribution of distance of CCC pattern of a certain cluster with the rest of the clusters. Cell-cell communication patterns are robust to subsampling within each hierarchical ligands/receptors group.
(e) Validation of CrossChatT to detect simulated trees. CrossChatT is capable of detecting all simulated trees in 1000 trials of simulation.

3.2.7 Robustness of CrossChatH

We tested the robustness of hierarchical clustering of CrossChatH to selection of K in K-nearest neighbor graph construction, noise, sparsity, and normalization, on a PBMC dataset (Figure B.1). To evaluate the robustness of hierarchical clustering with respect to K value in KNN graph construction, we ran the method with varying K values: 5, 10, 15, 20, 25, and 50. We calculate the multiscale similarity between clustering results generated by the different K values, and observed that the score between each multiscale clustering is almost always greater than 0.9, indicating a high robustness to K value selection.

To evaluate the robustness of clusters to noise, we generated three scaled Gaussian noise and added them to the gene expression matrix. Specifically, the Gaussian noise is multiplied by 3 different scales: 10^{-3}, 10^{-4}, and 10^{-5}. The multiscale similarity between the clustering with added Gaussian noise and the clustering with no noise is still very high, indicating a high robustness of our clustering to noise.

To evaluate the robustness to dropout, we chose four dropout rates: 0.1, 0.2, 0.3, and 0.5. The multiscale similarity decreases when dropout rate increases. Finally, to evaluate the robustness to different normalization methods, we tested three popular normalizations, including log-normalization, counts per million, and centered log-ratio transformation. We found that

log-normalization-based and counts-per-million-based normalization produced similar multiscale clustering results, and while normalization using centered log-ratio transformation produced somewhat different results.

Evaluating multiscale similarity: For two multiscale clustering results, $X = \{X_1, X_2, \cdots, X_m\}$ and $Y = \{Y_1, Y_2, \cdots, Y_n\}$, where X_i denotes the communities detected by the first clustering results at scale i, and Y_j has an analogous definition. The multiscale-similarity between X and Y is defined as:

$$\text{MSSim}(X, Y) = \frac{1}{2}\left[\frac{1}{m}\sum_{i=1}^{m}\min_{1\le j\le n}\text{ARI}(X_i, Y_j) + \frac{1}{n}\sum_{j=1}^{n}\min_{1\le i\le m}\text{ARI}(X_i, Y_j)\right]$$

where $\text{ARI}(X_i, Y_j)$ is the Adjusted Rand Index between two clustering results X_i and Y_j.

3.3 Discussion

A large number of tools are recently developed to infer cell-cell communications using either scRNA-seq or spatial transcriptomics datasets9–24. However, there are two major limitations of CellChat and other existing CCC inference methods: 1. They only infer CCC between predefined cell groups, neglecting the structure of cells determined by only ligands and receptors in CCC. 2. They calculate CCC between cell groups at a single "scale" of clusters, neglecting the multiple scales of cell state heterogeneity, i.e., there are substates within cell states, and substates within those.

Hierarchical structures within CCC can be defined from two complementary perspectives. First, from a global perspective, where there is hierarchy inherent to the analyzed cell population, which can naturally emerge, for example, due to differentiation, resulting in cell states with specialized characteristics. Ligand-receptor interactions may occur between cell groups at different "scales" of this hierarch, and resulting in a global hierarchy of CCC. Second, from a local perspective, where a group of cells that secrete a signal ligand (or signal receptor) may contain a subset of cells secreting a second ligand (or receptor). If we analyze

ligands and receptors with respect to their secreting cell groups, another hierarchy may be found. These local hierarchies may indicate a hierarchy of specialized functions.

Thus, we propose CrossChat, which detects and analyzes both global and local hierarchical structures within CCC determined by ligands/receptors using single-cell and spatial transcriptomics. CrossChat consists of two core methods:

1. CrossChatH detects a global hierarchical structure of cells based on gene expression of either all genes or ligands/receptors, using a multiscale community detection algorithm. We then infer hierarchical CCC patterns from these hierarchical clusters. 2. CrossChatT detects multiple local hierarchical structures directly within ligands/receptors, based on the cells that express these genes.

Both methods contain a variant that can be used on spatial transcriptomics datasets, by incorporating spatial information in hierarchy detection.

In general, CrossChat is particularly advantageous over existing CCC detection methods, like CellChat or CellPhoneDB in the following scenarios. First, when one wants to analyze CCC between cell groups at multiple clustering resolutions, rather than one single scale. Second, when one wants to investigate structures induced by cell-cell communication that are specific to signal ligands/receptors and independent of predefined cell type annotations. We believe CrossChat presents several major advances in the field: 1. It detects structures within CCC that do not rely on predefined cell type annotations. 2. It detects the structure of CCC determined by signal ligand expression or signal receptor expression, thus fits better with the real structure within CCC. 3. It provides a comprehensive view on hierarchical structures within CCC from both global and local perspectives.

Chapter 4

Applications of CrossChatH and CrossChatT to Single-cell and Spatial Transcriptomics Data

4.1 Introduction

Cell-cell communication (CCC) occurs across different biological scales, ranging from interactions between large groups of cells to interactions between individual cells, forming a hierarchical structure. Globally, CCC may exist between clusters or only subgroups of a cluster with varying size, while locally, a group of cells as sender or receiver may exhibit distinct signaling properties. Current existing methods infer CCC from single-cell RNA-seq or Spatial Transcriptomics only between predefined cell groups, neglecting the existing hierarchical structure within CCC that are determined by signaling molecules, in particular, ligands and receptors. In Chapter 3, we developed CrossChat, a novel computational framework designed to infer and analyze the hierarchical cell-cell communication structures using two complementary approaches: a global hierarchical structure using a multiresolution clustering method and multiple local hierarchical structures using a tree detection method.

This framework provides a comprehensive approach to understand the hierarchical relationships within CCC that govern complex tissue functions. In this chapter, by applying our method to two nonspatial scRNA-seq datasets sampled from COVID-19 patients and mouse embryonic skin, and two spatial transcriptomics datasets generated from Stereo-seq of mouse embryo and 10x Visium of mouse wounded skin, we showcase CrossChat's functionalities for analyzing structures within cell-cell communication for both scRNA-seq datasets and spatial transcriptomics datasets. The applications demonstrate that CrossChat can consistently detect biologically meaningful structures within CCC. In addition, we compared CrossChat with existing CCC inference tools to show the novelty of our method: 1. It detects structures within CCC that do not rely on predefined cell type annotations. 2. It detects the structure of CCC determined by signal ligand expression or signal receptor expression, thus fits better with the real structure within CCC. 3. It provides a comprehensive view on hierarchical structures within CCC from both global and local perspectives.

4.2 Results

4.2.1 CrossChatH identifies hierarchical clusters and specific interactions in COVID-19 patients

We demonstrate the functionalities of CrossChatH on a scRNA-seq dataset of COVID-19 patients. This dataset is clustered into eight cell types, including B cells, plasma B cells, CD4 cells, CD8 cells, NK cells, dendritic cells (DC), monocytes, and megakaryocytes (Figure 4.1a). Multi-resolution clustering shows that there are three clustering scales (Figure 4.1a). At each cluster, we assigned each cluster to a cell type based on their cell type composition (Figure C.1a). At scale 1, the coarsest scale, there are two clusters, where one contains a mixture of B cells and T cells, two major types of lymphocytes, and the other cluster contains monocytes. At scale 2, CrossChatH further clusters the lymphocyte cluster into B cells and two groups of T cells. At scale 3, the finest scale, finer subclusters are detected. Monocytes are further

clustered into two subclusters. B cells are clustered into three subclusters, identifying plasma B cells as a cell subtype. T cells are clustered into three subclusters, where CD4 and CD8 cells are identified cell subtypes. We note that megakaryocytes are merged into the T cell clusters. This is because megakaryocytes are scarce in number, and the community detection algorithm may not be robust at identifying small-number clusters.

CrossChatH detected 102 specific ligand-receptor pairs in this dataset (the top 15 pairs are presented in Figure C.1b). These identified CCCs generate a multi-resolution nature. For example, the ligand-receptor interaction CCL3-CCR1 interaction only occurs among a subset of monocytes, mono2 cells, and monocytes (Figure 4.1b). Interactions via RETN-TLR4 only occur between monocytes and another subset of monocytes, mono1 (Figure 4.1b). The interactions between the identified clusters account for most major interactions in the dataset (Figure 4.1b). We also use CrossChatH to visualize CCC strength between clusters at multiple scales (Figure 4.1c). These findings are consistent with previous studies, and further provide new insights into COVID-19. For example, one study showed that CCL3 is elevated in patients with more severe COVID-19 symptoms in monocytes sampled from PBMCs34. We observe that the specificity of CCL3 expression further suggests that CCL3 may be released by only a subset of monocytes in patients with severe COVID-19. Also, it has been shown that targeting CCL3-CCR1 may suppress "hyper activation" of the immune system in patients with COVID-19[28]. We found that CCL3 ligand and CCR1 receptor are only expressed in specific cell subtypes, suggesting that treatments should target a specific subset of monocytes as senders of CCL3 and all monocytes as receivers to suppress immune hyperactivation. A previous study also showed that activation of TLR4 is closely related to inflammation and immune hyperactivation, suggesting that targeting TLR4 could effectively reduce the inflammation in severe COVID-19 patients[69]. As we found that TLR4 is most specific to mono1 cells, this suggests that treatments should specifically target a subset of monocytes as a treatment to inflammation.

Furthermore, CrossChatH can quantify the similarity of ligand-receptor interactions or path-

ways based on their ligand (receptor) gene distribution over hierarchical clusters, and then group them based on their similarity (See Methods). Hierarchically grouping the top 15 specific ligand-receptor pairs yielded three clusters (Figure 4.1d). Group 1 represents interactions within T cells, including CCL5-ACKR2, CCL5-CCR4, GZMA-F2R (Figure C.1b). Group 2 represents interactions from monocytes to B cells, including LGALS9-IGHM and APP-CD74. Group 3 represents interactions whose receivers are monocytes, including CCL5-CCR1, GRN-SORT1, ANXA1-FPR1.

In addition to analyzing interaction patterns based on hierarchical clusters found through all genes, CrossChatH also allows to analyze hierarchical interactions between clusters found with respect to signal ligands or signal receptors only. By neglecting the influence of genes unrelated to ligand-receptor interactions, we generate hierarchical clusters that are potentially more aligned with distributions of ligands or receptors. Applying CrossChatH clustering on all ligands and all receptors respectively gives different hierarchical clusters (Figure C.1c–d). We assign cluster names based on their marker ligands/receptors. By observing the clustering results, the clusters based on all ligands are very similar to hierarchical clusters based on all genes, while the clusters found using all receptors are less aligned (Figure C.1c–d). CrosschatH provides two ways to visualize the pathway interactions between ligand and receptor trees: 1) by considering all original clusters with respect to ligands and receptors or 2) constructing a "pathway intersected" tree by focusing on a specific ligand cluster and all other cells (Figure 1.4e). As an example, we analyzed the Granulin pathway. While the interactions in Granulin pathway mostly occur between monocytes (PSAP is a marker ligand of this cluster), the intersected version shows that small portion of cells in T cells and B cells also secrete PSAP. Granulin has a central role in regulating the host response during infection. Its interaction with SORT1 can influence its levels and activity[59, 53, 12] and is hence seen as a promising therapeutic target for immune diseases. Our finding suggests that treatments should target monocytes. Furthermore, we clustered pathways based on their ligands and receptors distribution over hierarchical ligand clusters and receptor clusters (See

Methods). Applying hierarchical grouping yields three pathway clusters (Figure 4.1f). Pathways in group 1 are associated with CD8T and megakaryocytes cells and their subsets as signal senders, including CXCL, Insullin-like growth factor 2 (IFN-II), Protease-activated receptors (PARs), TGFb, and others. Pathways in group 2 are associated with all other T cells and B cells as signal senders, including Lymphotoxin (LT), CD45, Macrophage migration inhibitory factor (MIF), B-cell activating factor (BAFF), and others. Pathways in group 3 are associated with interactions involving monocytes, including Granulin (GRN), Synuclein (SN), RESISTIN, VISFATIN, and others. Collectively, CrossChatH detects hierarchical structures of cell states based on either all genes, or ligands/receptors, and identifies specific ligand-receptor interactions in hierarchical cell structures of a given scRNA-seq dataset.

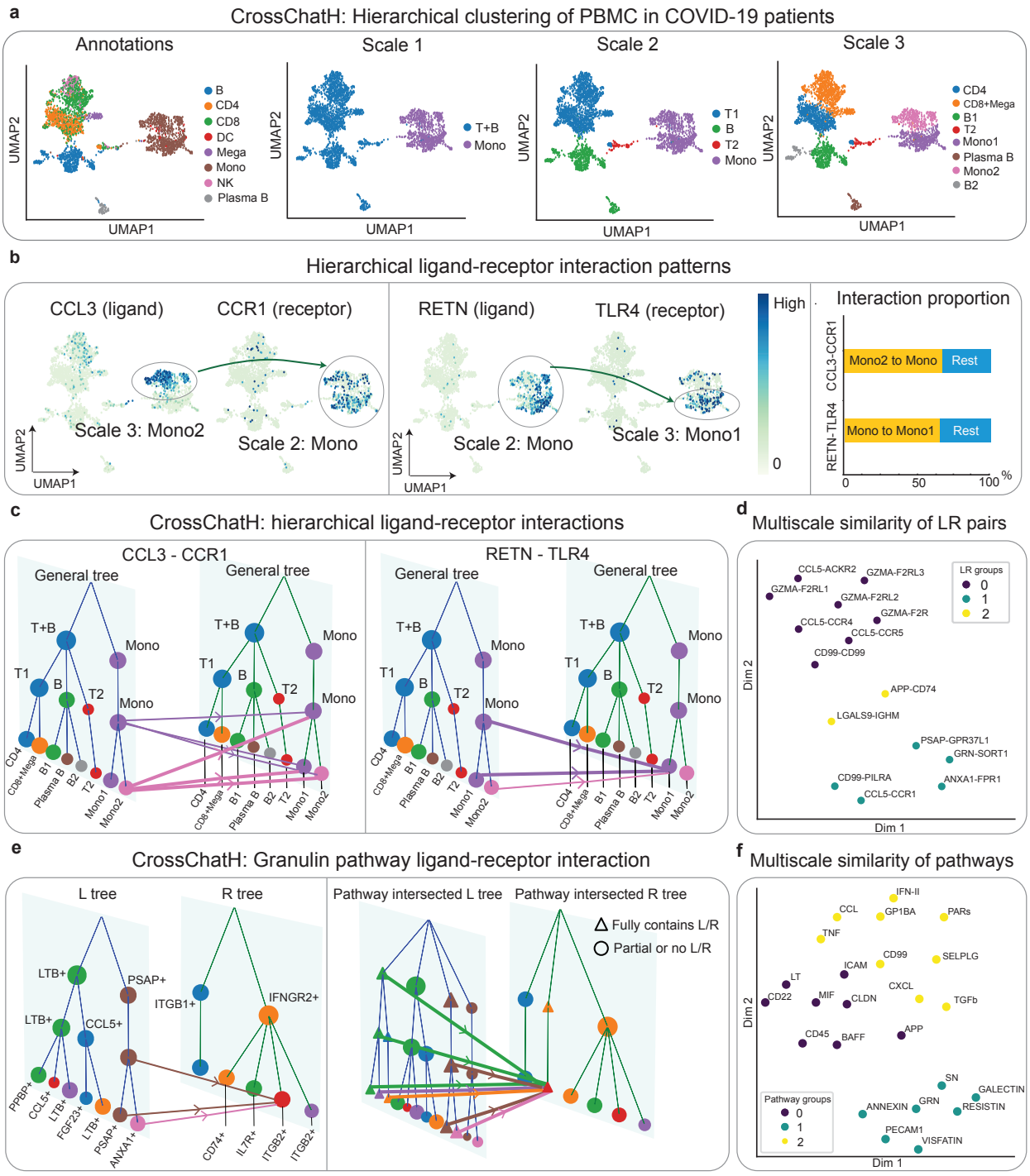


Figure 4.1: **COVID-19 PBMC analysis using CrossChatH.** See next page for figure description.

Figure 4.1 (continued). (a) Annotations and CrossChatH hierarchical clustering results of a scRNA-seq dataset of PBMC cells from COVID-19 patients. Hierarchical clustering detects three scales of clusters, and clusters at each scale align with certain biological clusters.
(b) Two ligand-receptor interactions that are specific to hierarchical clusters. CCL3 ligand is specific to Mono2, a subset of Mono, and CCR1 receptor is specific to Mono. RETN ligand is specific to Mono, and TLR4 receptor is specific to Mono1, another subset of Mono. Interaction proportion shows that the assigned cluster pairs account for most interactions through the CCL3–CCR1 and RETN–TLR4 interaction pairs.
(c) Hierarchical visualization of the two ligand-receptor interactions where ligand (receptor) clusters are from CrossChatH clustering based on ligands (receptors) only.
(d) Applying UMAP to the top 15 specific ligand-receptor pairs based on the similarity of ligand/receptor distribution over hierarchical clusters to visualize the interaction similarity in 2D.
(e) Hierarchical visualization of cell-cell communications of the Granulin pathway. The left side shows the visualization of interactions between hierarchical clusters generated based on ligands (receptors) only. The right side uses the support of ligands (receptors) to calculate interactions with each hierarchical cluster, so each hierarchical cluster expresses a subset of cells where every cell expresses the ligands (receptors), and the other subset expresses none.
(f) Applying UMAP to signaling pathways based on the similarity of ligands/receptors distribution over hierarchical clusters to visualize the interaction similarity in 2D. Source data are provided as a Source Data file.

4.2.2 CrossChatT identifies local hierarchical signaling structures driven by ligand-receptor interactions in mouse embryonic skin

While CrossChatH can analyze global hierarchical patterns in CCC, CrossChatT can provide a more specialized "local" perspective to analyze hierarchical structures within ligand-receptor interactions with respect to ligand or receptor gene expression. For example, in a mouse embryonic skin data, cells expresing ligand Ocln are a subset of cells expressing ligand Igf2, and cells expressing Itga6+Itgb4 (receptor associated with Ocln) are disjoint with Ocln, which is the receptor of itself (Figure 4.2a). The inclusive or disjoint relationship implicate hierarchical functions of cell groups, and such relationship of cells under CCC give rise to tree structures of ligands, receptors, and ligand-receptor interactions.

Applying CrossChatT on mouse embryonic skin E14 data[44], we found 142 hierarchical structures within ligands, and 11 hierarchical structures within receptors (Figure 4.2b). There are 58 ligands in total across the 142 ligand tree structures, and 18 receptors across all receptor tree structures, indicating many overlaps between these ligand/receptor trees. There are significantly more hierarchical structures among ligands than among receptors, indicating a more structured function of cells in sending ligands during mouse skin development. Many ligand trees begin with Igf2 as the root node. The signal ligand Igf2 is indeed expressed in most of the cells (Figure 4.2a). Igf2 is known to primarily regulate mitogenic functions and play an important role in cell growth and development[10]. Cells expressing Bmp5, Inhbb, Cdh1, Cdh5 form mutually disjoint groups (Figure C.1a-b). Bmp5 is mostly expressed in a subset of fibroblasts (FIB-A) and pericytes (Pericyte). In embryonic mouse skin, Bmp5 is involved in the control of epidermal homeostasis, growth of hair follicles, and melanogenesis[11]. Inhbb is mainly expressed in fibroblasts (Figure C.1a-b). It is a member of the TGF superfamily, and regulates the development of the skin[15]. Cdh1 is abundant in basal (Basal and Basal-P) and spinous cells (Spinous), which are both found in the epidermis of the skin. This gene is crucial for maintaining the structure of the epidermis by ensuring that cells adhere to each other to form a functional barrier[51]. Cdh5 is mainly expressed in endothelial cells (ENDO) and a minor subset of FIB-A cells (Figure C.1a-b). It ensures the integrity of blood vessels, which is important to the supply of nutrients and oxygen to the skin and removal of waste products[83]. Therefore, these signal-expressing groups, which are form disjoint subsets of the cells expressing Igf2, may be performing distinct functions in mouse embryonic skin. Furthermore, Ocln is mostly expressed in spinous cells (Spinous), a subcluster of the cell group expressing Cdh1, and is known to play a role in the formation and regulation of tight junctions[82, 70]. Ocln and Cdh1 are both important to epidermis structure, where functions of Ocln are more specific to spinous cells. Analyzing the hierarchical relationship of ligands/receptors in this way provides a deeper and more structured understanding of relations between different signaling functions of cell groups.

We then investigated the interactions between ligands hierarchies and receptors hierarchies (Figure 4.2c). In interaction tree 1, the adhesion gene, Ocln, is expressed by spinous cells, a group of differentiated keratinocytes. Ocln-Ocln interactions between spinous cells help form tight junctions. The ligand Igf2 is more widely expressed, and its interaction with Itga6_Itgb4 receptor, which is expressed in basal cells, may help promote their continued proliferation. In interaction tree 2, there exists a pair of hierarchical interactions within the WNT signaling pathway, which is pivotal to skin developmental processes including cell proliferation and differentiation[80]. Wnt4 is expressed widely by epidermal basal cells and spinous cells, as well as dermis-resident dendritic cells and fibroblasts, while Wnt3a is mainly expressed by basal cells. Since Wnt3a is known to inhibit cell proliferation in keratinocytes, while Wnt4 promotes cell proliferation in general[71], we may infer that Wnt3a ligand binding competes with the binding of Wnt4, regulating cell proliferation in basal cells.

CrossChatT can analyze multiple ligand trees (or receptor trees) in tandem (Figure 4.2d), facilitating understanding of relationship within ligands (receptors). We note that these "aggregated" ligand trees (or receptor trees) do not necessarily exhibit the disjoint relationships between ligands in the individual trees. To measure the importance of ligands/receptors in their hierarchical structures, CrossChatT measures frequency of the ligand/receptor occurrence across all ligands/receptors trees (Figure 4.2e). Igf2 is ranked first due to its wide expression by cells. While Ptprf is ranked first among receptors due to its wide expression, Ocln, which is more specifically expressed in spinous cells, is also ranked highly as its expression is disjoint from many other receptors. In addition to detecting hierarchical ligands/receptors trees, CrossChatT is also capable of detecting trees within cells that are involved in the ligand-receptor interactions (Figure 4.2f). Tree 1 is associated with the WNT signaling pathway, while Tree 2 is associated with the MK signaling pathway.

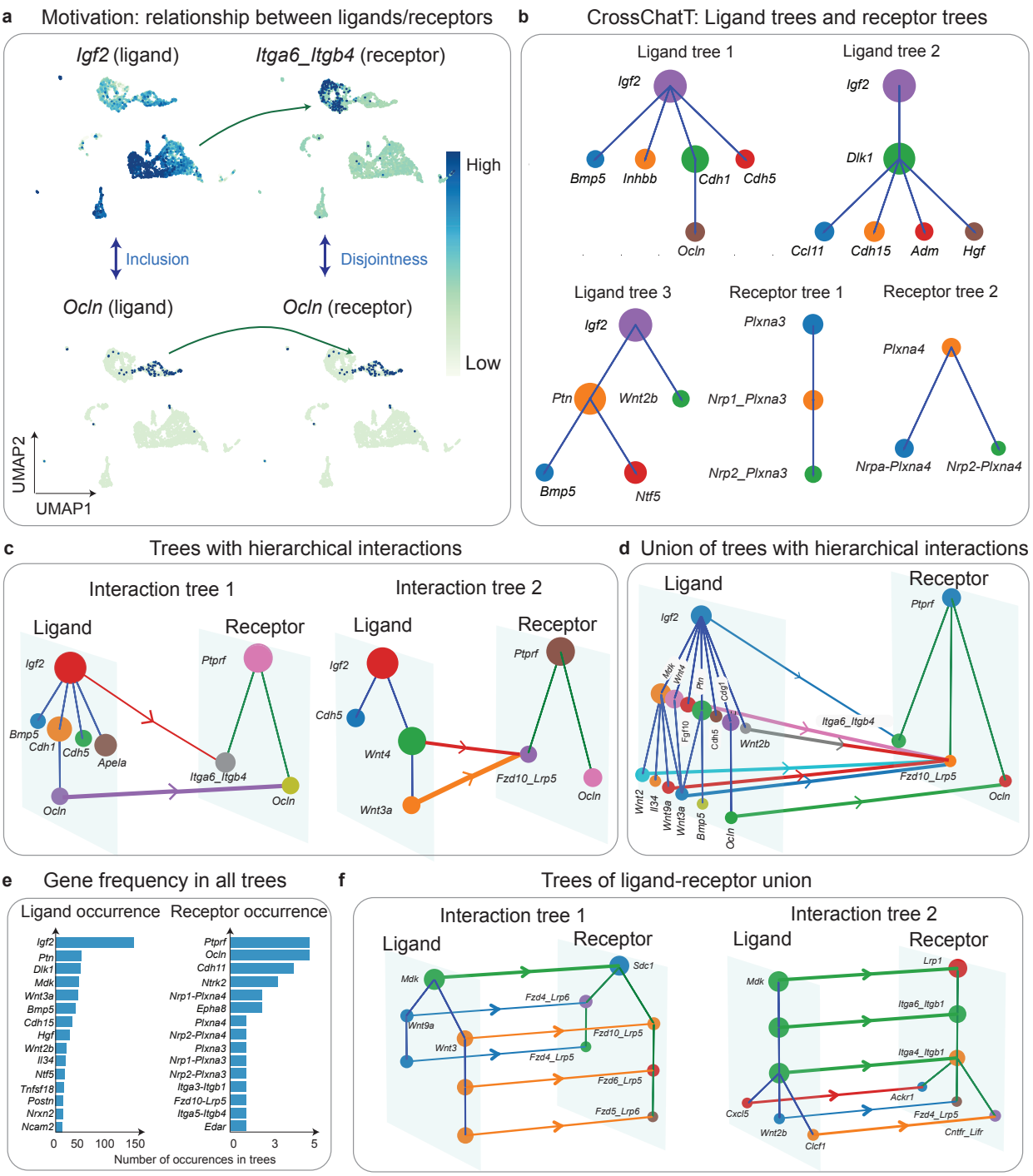


Figure 4.2: **Applications of CrossChatT to mouse skin cells during wound healing.** See next page for detailed caption.

Figure 4.2 (continued). (a) An example of hierarchical structures of ligand-receptor interactions. Igf2 and Ocln are inclusive as ligands. Itga6_Itgb4 and Ocln are disjoint as receptors. (b) Examples of ligand tree structures and receptor tree structures detected by CrossChatT. (c) Two examples of hierarchical interactions occurring between a ligand tree and a receptor tree. (d) Hierarchical interactions between the union of ligand trees and the union of receptor trees. (e) Frequency of ligands/receptors occurrence across all ligand/receptor trees. (f) Examples of tree structures of ligand-receptor unions detected by CrossChat. Source data are provided as a Source Data file.

4.2.3 CrossChatS reveals hierarchical clusters and ligand-receptor interactions in spatial datasets

CrossChat can also be applied to spatial transcriptomics data by incorporating spatial information in both CrossChatH and CrossChatT. We demonstrate utility of CrossChatH-S to a Stereo-seq data of a mouse embryo on day E16.5[20]. We compared the hierarchical clustering results between two trials: one trial use spatial information, and the other trial does not use spatial information (Figure C.2a-b). We can see that when we incorporate spatial locations by concatenating the spatial vectors to PCA embeddings, the clusters detected tend to be more spatially coherent. For example, the neurons are further segmented into multiple spatial regions. We also observed that the multiscale clusters generated with spatial information have higher spatial neighborhood enrichment scores, indicating higher spatial coherence (Figure C.2a-b). Using spatial information (see Method), CrossChatH identifies three scales of clustering (Figure 4.3). At scale 3, the clusters correspond to biologically meaningful clusters (Figure C.3). With the addition of spatial information, CrossChat also groups spatially adjacent spots. For example, MSCN (mid/hindbrain and spinal cord neuron) is partitioned into two subclusters at scale 3, where subcluster 1 corresponds to neurons in the brain, and subcluster 2 corresponds to spinal cord neurons. Next, we investigated specific ligand-receptor interactions between clusters (Figure C.3). Igf2-Igf2r interaction mainly occurs within the mesoderm (Figure 4.3b-c). This interaction is crucial to organ development in mouse embryo, and it occurs at a large scale[63]. Interactions between synaptic adhesion molecules

Nlgn1 and Nrxn1, which known to be crucial to synapse formation and maturation[113, 29], are specifically expressed in neurons in the brain and spinal cord. Hierarchical clustering is also performed using only ligands/receptors on this dataset (Figure C.3c-d), but hierarchical patterns of ligands/receptors were not identified. This indicates that during mouse embryo early development, where most cells have not fully matured, ligand/receptor interactions are not specialized to a particular cell group. Next, by hierarchically grouping active pathways, we detected two pathway groups (Figure 4.3d). The first pathway group is associated with signaling mainly sending from brain and spinal cord neurons (MSCN), including the pathways NCAM (Neural cell adhesion molecule), NRG (Neuregulin), and NRXN (Neurexin) (Figure C.4a). The second pathway group is associated with interactions within all other spots (Figure C.4b). Hierarchically grouping the top 20 specific ligand-receptor pairs yields three clusters (Figure C.4c). Group 1 includes interactions in MDK (Midkine) pathway and IGF (Insulin-like growth factor) pathway, which may both contribute to growth of the mouse embryo. Group 2 mainly contains interactions from Dlk1 to Notch receptors, including Notch1, Notch2, and Notch3. Dlk1 has been shown to inhibit Notch signaling[37]. Group 3 contains signaling between neurons, notably Nrxn1-Nlgn3 and Nrg3-Erbb4.

CrossChatT-S uses spatial information to restrict the spatial distance between ligands and receptors when detect ligand/receptor trees (see Methods). Applied to a Visium dataset of wounded mouse skin sampled at post-operative day 7[40], we find many hierarchical structures within ligands and receptors during wound healing, including 124 ligand trees and 456 receptor trees (Figure C.5a). There are 58 and 141 distinct ligands and receptors, respectively, across all ligands and receptors trees (igure 4.3e). By measuring the occurrence of each ligand and receptor (Figure C.5b), we see that C3, Postn, Ccl8 are most frequent across hierarchical structures in ligands, while Cd74, Cd74_Cd44, Cd44 are highly frequent across receptor trees. The ligands C3, Postn, and Ccl8 are expressed broadly by dermis-resident spots. While the Cd74 receptor is almost expressed by all spots, expression of Cd74_Cd44 and Cd44 receptors is concentrated around the wounded regions (Figure C.5ac).

We then investigated interactions between ligand trees and receptor trees (igure 4.3f). In interaction tree 1, Periostin (Postn), a pro-fibrogenic secreted glycoprotein, interacts with its two receptors. Periostin is known to promote injury closure by facilitating the activation, differentiation, and contraction of fibroblasts[26]. Thy1 ligand, which is expressed by a subset of Postn-expressing cells, affects cell adhesion and fibroblasts proliferation and migration, suggesting a more specific function of these cells during tissue repair[97]. In interaction tree 2, Ccl8 and Ccl25 both interact with Ackr4, while Ccl25 is only expressed in a subset of Ccl8-expressing cells. Chemokines are major regulators of the wound-healing process57. Ccl8 is known to recruit inflammatory cells including neutrophils and macrophages, which release growth factors and cytokines to improve wound healing[79]. One study found that Ccl25 is enriched during oral wound healing process that facilitates leukocyte recruitment, which helps to protect against infection and promote wound healing[96, 36]. Our finding suggests that there is a hierarchical relationship among chemokines which may facilitate specialized roles during skin wound healing.

Comparison between CrossChat and other CCC detection methods To show the differences between existing methods and CrossChat, we compared CrossChat with two representative CCC methods: CellChat9 for nonspatial mouse embryonic skin scRNA-seq, and COMMOT22 for spatial mouse embryo data. We also use NeST[99], a spatial multiscale spot detection method, to produce spatial clusters for CCC detection. By comparing CrossChatH's output for the ligand-receptor interaction, CCL3-CCR1, to CellChat, we can see that CellChat only shows that the interaction occurs between monocytes. But CrossChatH is more precise and finds that CCL3–CCR1 signaling is mainly sent from a specific subset of monocytes, Mono2 (Figure C.6ab). Also, by comparing CrossChatT's output with CellChat, we can see CrossChatT is able to detect the inclusion relationship between the two interactions: Igf2-Itga6_Itgb4, and Ocln-Ocln. Specifically, the cells expressing the Igf2 ligand also express Ocln ligand, and the cells expressing Itga6_Itgb4 receptor also express Ocln receptor. Such hierarchical relationships among ligand-receptor interactions cannot be

identified by CellChat and existing CCC methods (Figure C.6c-e).

Furthermore, we compared the output of CrossChatH with the output of the spatial CCC method COMMOT on spatial clusters produced by NeST (Figure C.7a-e). While NeST clustering generates multiple sparse and tiny clusters, CrossChatH detects a more general hierarchical relationship among cells, where each scale in the clustering covers all cells in the dataset (Figure C.7c-e). By comparing CrossChatT to COMMOT where clusters are assigned by biological annotations, we again see that CrossChatT detects the structural relationship between the interactions, which cannot be detected by COMMOT. For example, the Ccl25-Ackr4 interaction is only involved in cells which interact through Ccl8-Ackr4 (Figure C.7f–h).

Overall, by comparing with existing CCC detection methods, we can see that CrossChat (both CrossChatH and CrossChatT) is, in principle, different from existing methods in terms of clustering choice. CrossChat aims to detect the hierarchy of cells based on CCC relations. However existing CCC methods rely on predefined clustering and calculate the interactions between these predefined clusters. Those predefined clusters are produced by standard clustering results like Louvain community detection, which fails to consider hierarchical CCC relationships.

We also tested how different CCC calculation methods may affect the detected hierarchical structures. Using the COVID-19 scRNA-seq dataset, we recalculated CCC using NATMI[49], LIANA[32], SingleCellSignalR[16] (Figure C.8a–d). Using the mouse skin wound 10x Visium data, we recalculated CCC using Giotto[34], SpatialDM[61] and spaCI[93] (Figure C.9a–d). We can see that the hierarchical cell structures identified remain consistent across various CCC calculation methods. However, interaction strengths exhibit minor variations due to the different scoring functions employed by each method. This consistency underscores a key aspect of CrossChat's functionality: it initially identifies hierarchical structures based on only on ligand/receptor expression, independent of the CCC calculation method used. Thus, while changes in CCC calculation methods influence the inferred interaction strengths, they

do not alter the underlying detected structures.

For further validation, we performed hierarchical clustering of CCC calculated using single-cell-level CCC methods, including NICHES and Scriabin on COVID-19 patients (Figure C.10ab). Many scattered and small clusters are detected, and they do not have hierarchical structures. This may be due to the limitations of current single-cell-level CCC methods, which do not accurately capture cell-level interactions and may introduce noise in their inference. Moreover, we attempted to explore the clustering of cell pairs using traditional hierarchical clustering and visualize the clusters using dendrograms. Specifically, we used NICHES to calculate single-cell level CCC and then performed traditional hierarchical clustering using dendrograms on cell pairs (Figure C.10c). However, we observed one major limitation in this type of approach: although our test dataset contains only 3000 cells, there are nine million possible cell pairs, making hierarchical clustering computationally infeasible due to both memory and complexity constraints. To gain at least a basic understanding of how traditional hierarchical clustering might work on cell pairs, we reduced the dataset to 100 cells, resulting in 10,000 cell pairs—small enough to meet computational constraints. We generated a heatmap that illustrates the overlap of sender-receiver cell groups across 20 clusters of cell pairs. However, the cell-pair structure is much less interpretable than the original cell groups, largely because different clusters of cell pairs can include identical sets of cells (simply paired differently), making it difficult to visually distinguish meaningful differences between the clusters (Figure C.10d). For example, the Jaccard similarity of sender/receiver cells in cell pair cluster 5 and 9 is 0.98, indicating that the group of senders and receivers are indeed almost identical even though they're in two different clusters of cell-pairs. To ensure the robustness of this finding, we also performed subsampling for 20 times. We found that the average number of pairs with similarity higher than 0.5 is 11.6, indicating that many of the clusters of cell pairs have a high overlap.

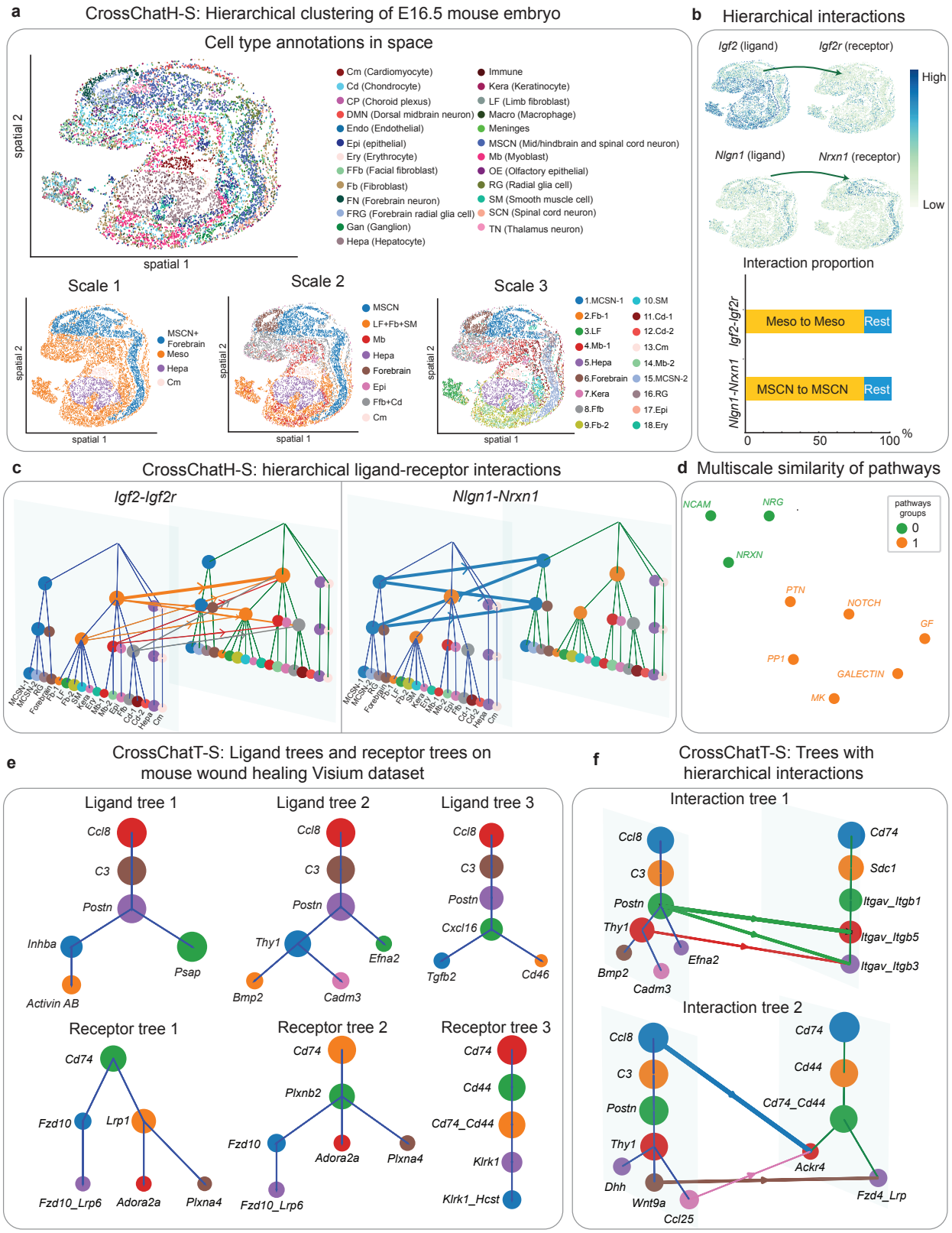


Figure 4.3: **Applications of CrossChatS to spatial datasets.** See next page for detailed caption.

Figure 4.3 (continued). (a) Annotations and CrossChatH-S hierarchical clustering results of a Stereo-seq dataset of mouse embryo at E16.5. Spatial hierarchical clustering detects three scales of clusters, which are shown in spatial coordinates. (b) Two ligand-receptor interactions that are specific to spatial hierarchical clusters. Ccl3 ligand is specific to Mono2, a subset of Mono, and Ccr1 receptor is specific to Mono. Retn ligand is specific to Mono, and Tlr4 receptor is specific to Mono1, another subset of Mono. Interaction proportion shows that the assigned cluster pairs contain the majority of interactions for both Ccl3–Ccr1 and Retn–Tlr4 interactions. (c) Hierarchical visualization of the two ligand-receptor interactions where ligand (receptor) clusters are from CrossChatH-S clustering based on ligands (receptors) only. (d) Applying UMAP to pathways based on the similarity of ligand/receptor distribution over spatial hierarchical clusters to visualize the interaction similarity in 2D. (e) Examples of ligand tree structures and receptor tree structures detected by CrossChatT-S. (f) Two examples of hierarchical interactions in space occurring between a ligand tree and a receptor tree. Source data are provided as a Source Data file.

4.2.4 Comparison between CrossChat and other CCC detection methods

To show the differences between existing methods and CrossChat, we compared CrossChat with two representative CCC methods: CellChat61 for nonspatial mouse embryonic skin scRNA-seq, and COMMOT62 for spatial mouse embryo data. We also use NeST63, a spatial multiscale spot detection method, to produce spatial clusters for CCC detection. By comparing CrossChatH's output for the ligand-receptor interaction, CCL3-CCR1, to CellChat, we can see that CellChat only shows that the interaction occurs between monocytes. But CrossChatH is more precise and finds that CCL3–CCR1 signaling is mainly sent from a specific subset of monocytes, Mono2 (Figure C.6ab). Also, by comparing CrossChatT's output with CellChat, we can see CrossChatT is able to detect the inclusion relationship between the two interactions: Igf2-Itga6_Itgb4, and Ocln-Ocln. Specifically, the cells expressing the Igf2 ligand also express Ocln ligand, and the cells expressing Itga6_Itgb4 receptor also express Ocln receptor. Such hierarchical relationship among ligand-receptor interactions cannot be identified by CellChat and existing CCC methods (Figure C.6c-e).

Furthermore, we compared the output of CrossChatH with the output of the spatial CCC

method COMMOT on spatial clusters produced by NeST (Figure C.7a-e). While NeST clustering generates multiple sparse and tiny clusters, CrossChatH detects a more general hierarchical relationship among cells, where each scale in the clustering covers all cells in the dataset (Figure C.7c-e). By comparing CrossChatT to COMMOT where clusters are assigned by biological annotations, we again see that CrossChatT detects the structural relationship between the interactions, which cannot be detected by COMMOT. For example, the Ccl25-Ackr4 interaction is only involved in cells which interact through Ccl8-Ackr4 (Figure C.7f-h).

Overall, by comparing with existing CCC detection methods, we can see that CrossChat (both CrossChatH and CrossChatT) is, in principle, different from existing methods in terms of clustering choice. CrossChat aims to detect the hierarchy of cells based on CCC relations. But existing CCC methods rely on predefined clustering and calculate the interactions between these predefined clusters. Those predefined clusters are produced by standard clustering results like Louvain community detection, which fails to consider hierarchical CCC relationships.

We also tested how different CCC calculation methods may affect the detected hierarchical structures. Using the COVID-19 scRNA-seq dataset, we recalculated CCC using NATMI64, LIANA65, SingleCellSignalR66,67 (Figure C.8a-d). Using the mouse skin wound 10x Visium data, we recalculated CCC using Giotto67, SpatialDM68 and spaCI69 (Figure C.9a-d). We can see that the hierarchical cell structures identified remain consistent across various CCC calculation methods. However, interaction strengths exhibit minor variations due to the different scoring functions employed by each method. This consistency underscores a key aspect of CrossChat's functionality: it initially identifies hierarchical structures based on only on ligand/receptor expression, independent of the CCC calculation method used. Thus, while changes in CCC calculation methods influence the inferred interaction strengths, they do not alter the underlying detected structures.

For further validation, we performed hierarchical clustering of CCC calculated using single-

cell-level CCC methods, including NICHES and Scriabin on COVID-19 patients (Figure C.10ab). Many scattered and small clusters are detected, and they do not have hierarchical structures. This may be due to the limitations of current single-cell-level CCC methods, which do not accurately capture cell-level interactions and may introduce noise in their inference. Moreover, we attempted to explore the clustering of cell pairs using traditional hierarchical clustering and visualize the clusters using dendrograms. Specifically, we used NICHES to calculate single-cell level CCC and then perform traditional hierarchical clustering using dendrograms on cell pairs (Figure C.10c). However, we observed one major limitation in this type of approach: although our test dataset contains only 3,000 cells, there are nine million possible cell pairs, making hierarchical clustering computationally infeasible due to both memory and complexity constraints. To gain at least a basic understanding of how traditional hierarchical clustering might work on cell pairs, we reduced the dataset to 100 cells, resulting in 10,000 cell pairs—small enough to meet computational constraints. We generated a heatmap that illustrates the overlap of sender-receiver cell groups across 20 clusters of cell pairs. However, the cell-pair structure is much less interpretable than the original cell groups, largely because different clusters of cell pairs can include identical sets of cells (simply paired differently), making it difficult to visually distinguish meaningful differences between the clusters (Figure C.10d). For example, the Jaccard similarity of sender/receiver cells in cell pair cluster 5 and 9 is 0.98, indicating that the group of senders and receivers are indeed almost identical even though they're in two different clusters of cell-pairs. To ensure the robustness of this finding, we also performed subsampling for 20 times. We found that the average number of pairs with similarity higher than 0.5 is 11.6, indicating that many of the clusters of cell pairs have a high overlap.

4.3 Discussion

A large number of tools are recently developed to infer cell-cell communications using either scRNA-seq or spatial transcriptomics datasets9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24. However, the major limitations of them are: 1. They infer CCC based on predefined cell groups, and overlook the cell structures that form with respect to ligands or receptors. 2. They overlook the existing clusters of cell states at different scales. To overcome these limitations, we have developed CrossChat to detect hierarchical structures within CCC from scRNA-seq and spatial transcriptomics data, which infers and analyzes hierarchical CCC structures from complementary global and local perspectives.

By applying CrossChat to two nonspatial scRNA-seq datasets, PBMC cells sampled from COVID-19 patients and mouse embryonic skin at day E14.5, and two spatial transcriptomics datasets, a Stereo-seq dataset of mouse embryo at day E16.5 and a Visium dataset of mouse wounded skin, CrossChat is always capable of detecting biologically meaningful hierarchical structures within CCC, and improve the understanding of CCC from a structured way. As an example, CrossChatH recovers the hierarchical structures of PBMC cells in COVID-19 patients, and identifies specific ligand-receptor interactions within monocytes, and within subsets of monocytes, providing deeper understanding of structured functions of cells. CrossChatT detects multiple tree structures within ligands/receptors in mouse embryonic skin and provides a structural understanding of ligands/receptors in terms of their inclusion and disjointness relationship. Notably, we observed that while the growth factor, Igf2, expressed by many cells, the ligands Bmp5, Inhbb, Cdh1, Cdh5 are expressed by cells that form mutually disjoint groups, possibly due to these cell groups performing different functions. We also observed that the adhesion gene, Ocln, is expressed only within a subset of Cdh5-expressing keratinocytes, specifically, spinous keratinocytes, suggesting a specialized role for differentiated keratinocytes.

Several directions for future studies on hierarchical patterns of CCC can be anticipated. First, while we have shown that CrossChat can be easily extended for spatial applications,

there may be other ways to utilize the spatial information. For example, one may restrict CrossChatT-S analysis of hierarchical structures to a local spatial region, rather than imposing that there is a hierarchy across all cells. Furthermore, as more time-series datasets of scRNA-seq and spatial transcriptomics become available [33, 86, 27], CrossChat may be applied to gain further understanding of temporal changes of hierarchical structures within CCC. Also, with the recent availability of 3D spatiotemporal data[101], CrossChat may be used to analyze hierarchical structures within CCC in 3D, with the potential of gaining more reliable understanding of tissues. We anticipate further studies of hierarchical structures within CCC will prosper as the sequencing technologies continue to develop rapidly.

Bibliography

[1] Armingol, e., baghdassarian, h. m. lewis, n. e. the diversification of methods for studying cell–cell interactions and communication. nat. rev. genet.25, 381–400 (2024).

[2] A. Aksakal, B. Kerget, F. Kerget, and S. Aikin. Evaluation of the relationship between macrophage migration inhibitory factor level and clinical course in patients with covid-19 pneumonia. *Journal of Medical Virology*, 93:6519–6524, 2021.

[3] A. A. Almet, Z. Cang, S. Jin, and Q. Nie. The landscape of cell–cell communication through single-cell transcriptomics. *Curr. Opin. Syst. Biol.*, 26, 2021.

[4] E. Armingol, A. Officer, O. Harismendy, and N. Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88, 2021.

[5] A. Arnaudon, D. J. Schindler, R. L. Peach, A. Gosztolai, M. Hodges, M. T. Schaub, and M. Barahona. Algorithm 1044: Pygenstability, a multiscale community detection framework with generalized markov stability. *ACM Transactions on Mathematical Software*, 50(2):1–8, 2024.

[6] D. Arnol, D. Schapiro, B. Bodenmiller, J. Saez-Rodriguez, and O. Stegle. Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. *Cell Reports*, 29(1):202–211.e6, 2019.

[7] P. Author and R. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation. *Journal of the Royal Statistical Society*, 85(1), 1922.

[8] G. Baruzzo, G. Cesaro, and B. di Camillo. Identify, quantify and characterize cellular communication from single-cell rna sequencing data with scseqcomm. *Bioinformatics*, 38(7):1920–1929, 2022.

[9] C. Bleilevens, J. Soppert, A. Hoffmann, T. Breuer, J. Bernhagen, L. Martin, et al. Macrophage migration inhibitory factor (mif) plasma concentration in critically ill covid-19 patients: A prospective observational study. *Diagnostics*, 11:332, 2021.

[10] A. J. Blyth, N. S. Kirk, and B. E. Forbes. Understanding igf-ii action through insights into receptor binding and activation. *Cells*, 9, 2020.

[11] V. A. Botchkarev. Bone morphogenetic proteins and their antagonists in skin and hair follicle biology. *J. Investigative Dermatol.*, 120, 2003.

[12] F. Brandes. Progranulin signaling in sepsis, community-acquired bacterial pneumonia and covid-19: a comparative, observational study. *Intensive Care Med. Exp.*, 9, 2021.

[13] C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph [h]. *Commun. ACM*, 16, 1973.

[14] R. Browaeys, W. Saelens, and Y. Saeys. Nichenet: modeling intercellular communication by linking ligands to target genes. *Nature Methods*, 17(2):159–162, 2020.

[15] C. W. Brown, L. Li, D. E. Houston-Hawkins, and M. M. Matzuk. Activins are critical modulators of growth and survival. *Mol. Endocrinol.*, 17, 2003.

[16] S. Cabello-Aguilar, M. Alame, F. Kon-Sun-Tack, C. Fau, M. Lacroix, and J. Colinge. Singlecellsignalr: inference of intercellular networks from single-cell transcriptomics. *Nucleic acids research*, 48(10):e55–e55, 2020.

[17] Z. Cang and Q. Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature Communications*, 11(1), 2020.

[18] Z. Cang, Y. Zhao, A. Almet, A. Stabell, R. Ramos, M. Plikus, S. Atwood, and Q. Nie. Screening cell–cell communication in spatial transcriptomics via collective optimal transport. *Nature Methods*, 2023.

[19] S. Chaurasia, M. Ekhlak, G. Kushwaha, V. Singh, R. Mallick, and D. Dash. Notch signaling functions in noncanonical juxtacrine manner in platelets to amplify thrombogenicity. *eLife*, 11:1–25, 2022.

[20] A. Chen. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185, 2022.

[21] J. Cheng, J. Zhang, Z. Wu, and X. Sun. Inferring microenvironmental regulation of gene expression from single-cell rna sequencing data using scmlnet with an application to covid-19. *Briefings in Bioinformatics*, 00:1–18, July 2020.

[22] J. Cheng, J. Zhang, Z. Wu, and X. Sun. Inferring microenvironmental regulation of gene expression from single-cell rna sequencing data using scmlnet with an application to covid-19. *Briefings in Bioinformatics*, 22(2):988–1005, 2021.

[23] H. Choi, J. Sheng, D. Gao, F. Li, A. Durrans, S. Ryu, S. Lee, and et al. Transcriptome analysis of individual stromal cell populations identifies stroma-tumor crosstalk in mouse lung cancer model. *Cell Reports*, 10(7):1187–1201, 2015.

[24] S. Choudhary and R. Satija. Comparison and evaluation of statistical error models for scrna-seq. *Genome Biology*, 23(1), 2022.

[25] A. Cillo, C. Kürten, T. Tabib, Z. Qi, S. Onkar, T. Wang, A. Liu, U. Duvvuri, S. Kim, R. Soose, S. Oesterreich, W. Chen, R. Lafyatis, T. Bruno, R. Ferris, and D. Vignali. Immune landscape of viral- and carcinogen-driven head and neck cancer. *Immunity*, 52(1):183–199.e9, 2020.

[26] S. J. Conway. The role of periostin in tissue remodeling across health and disease. *Cell. Mol. Life Sci.*, 71, 2014.

[27] D. P. Cook and B. C. Vanderhyden. Context specificity of the emt transcriptional response. *Nat. Commun.*, 11, 2020.

[28] F. Coperchini. The cytokine storm in covid-19: Further advances in our understanding the role of specific chemokines involved. *Cytokine Growth Factor Rev.*, 58, 2021.

[29] A. M. Craig and Y. Kang. Neurexin-neuroligin signaling in synapse development. *Curr. Opin. Neurobiol.*, 17, 2007.

[30] J. Cyster and C. Allen. B cell responses: Cell interaction dynamics and decisions. *Cell*, 177:524–540, 2019.

[31] H. Dheir, S. Yaylaci, S. Sipahi, A. C. Genc, D. Cekic, F. B. Tuncer, et al. Does macrophage migration inhibitory factor predict the prognosis of covid-19 disease? *Journal of Infection in Developing Countries*, 15:398–403, 2021.

[32] D. Dimitrov, D. Türei, M. Garrido-Rodriguez, P. Burmedi, J. Nagai, C. Boys, R. Ramirez Flores, H. Kim, B. Szalai, I. Costa, A. Valdeolivas, A. Dugourd, and J. Saez-Rodriguez. Comparison of methods and resources for cell-cell communication inference from single-cell rna-seq data. *Nature Communications*, 13(1):3224, 2022.

[33] J. Ding, N. Sharon, and Z. Bar-Joseph. Temporal modelling using single-cell transcriptomics. *Nature Reviews Genetics*, 23:355–368, 2022.

[34] R. Dries, Q. Zhu, R. Dong, C. Eng, H. Li, K. Liu, Y. Fu, and et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, 22(1), 2021.

[35] M. Efremova, M. Vento-Tormo, S. A. Teichmann, and R. Vento-Tormo. Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature protocols*, 15(4):1484–1506, 2020.

[36] S. A. Eming, T. Krieg, and J. M. Davidson. Inflammation in wound repair: Molecular and cellular mechanisms. *J. Investigative Dermatol.*, 127, 2007.

[37] F. A. Falix, D. C. Aronson, W. H. Lamers, and I. C. Gaemers. Possible roles of dlk1 in the notch pathway during development and disease. *Biochim Biophys. Acta Mol. Basis Dis.*, 1822, 2012.

[38] P. F. Ferreira, J. Kuipers, and N. Beerenwinkel. Identifying hierarchical cell states and gene signatures with deep exponential families for single-cell transcriptomics. *bioRxiv*, page 2022.10.15.512383, 2024. Preprint.

[39] M. Fischer, A. E. Schade, T. B. Branigan, G. A. Müller, and J. A. DeCaprio. Coordinating gene expression during the cell cycle. *Trends Biochemical Sci.*, 47, 2022.

[40] D. S. Foster. Integrated spatial multiomics reveals fibroblast fate during tissue repair. *Proc. Natl Acad. Sci. USA*, 118, 2021.

[41] D. L. Gibbs, B. Aguilar, V. Thorsson, A. V. Ratushny, and I. Shmulevich. Patient-specific cell communication networks associate with disease progression in cancer. *Frontiers in genetics*, page 1489, 2021.

[42] S. Gillespie and M. Monje. An active role for neurons in glioma progression: making sense of scherer's structures. *Neuro-Oncology*, 20(10):1292–1299, 2018.

[43] C. Guerrero-Juarez, G. Lee, Y. Liu, S. Wang, M. Karikomi, Y. Sha, R. Chow, and et al. Single-cell analysis of human basal cell carcinoma reveals novel regulators of tumor growth and the tumor microenvironment. *Science Advances*, 8(23):1–18, 2022.

[44] K. Gupta. Single-cell analysis reveals a hair follicle dermal niche molecular differentiation trajectory that begins prior to morphogenesis. *Dev. Cell*, 48, 2019.

[45] D. Hanahan and L. Coussens. Accessories to the crime: Functions of cells recruited to the tumor microenvironment. *Cancer Cell*, 21:309–322, 2012.

[46] D. Hanahan and R. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[47] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, S. Zheng, A. Butler, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184:3573–3587.e29, 2021.

[48] R. S. Hoekzema. Multiscale methods for signal selection in single-cell data. *Entropy*, 24, 2022.

[49] R. Hou, E. Denisenko, H. Ong, J. Ramilowski, and A. Forrest. Predicting cell-to-cell communication networks using natmi. *Nature Communications*, 11(1), 2020.

[50] Y. Hu, T. Peng, L. Gao, and K. Tan. Cytotalk: De novo construction of signal transduction networks using single-cell transcriptomic data. *Science Advances*, 7, 2021.

[51] X. Huang. Cdh1 is identified as a therapeutic target for skin regeneration after mechanical loading. *Int J. Biol. Sci.*, 17, 2021.

[52] J. Jakobsson, O. Spjuth, and M. Lagerström. scconnect: a method for exploratory analysis of cell–cell communication based on single-cell rna-sequencing data. *Bioinformatics*, 2021.

[53] J. Jian, J. Konopka, and C. Liu. Insights into the role of progranulin in immunity, infection, and inflammation. *J. Leukoc. Biol.*, 93, 2013.

[54] S. Jin, C. Guerrero-Juarez, L. Zhang, I. Chang, R. Ramos, C. Kuan, P. Myung, M. Plikus, and Q. Nie. Inference and analysis of cell-cell communication using cellchat. *Nature Communications*, 12(1):1–20, 2021.

[55] T. Johung and M. Monje. Neuronal activity in the glioma microenvironment. *Current Opinion in Neurobiology*, 47:156–161, 2017.

[56] T. Kawabe, M. Matsushima, N. Hashimoto, K. Imaizumi, and Y. Hasegawa. Cd40/cd40 ligand interactions in immune responses and pulmonary immunity. *Nagoya Journal of Medical Science*, 73:69–78, 2011. Invited Review Article.

[57] M. Kuchroo. Multiscale phate identifies multimodal signatures of covid-19. *Nat. Biotechnol.*, 40, 2022.

[58] M. Kumar, J. Du, G. Lagoudas, Y. Jiao, A. Sawyer, D. Drummond, D. Lauffenburger, and A. Raue. Analysis of single-cell rna-seq identifies cell-cell communication associated with tumor characteristics. *Cell Reports*, 25(6):1458–1468.e4, 2018.

[59] Y. J. Lan, N. B. Sam, M. H. Cheng, H. F. Pan, and J. Gao. Progranulin as a potential therapeutic target in immune-mediated diseases. *J. Inflamm. Res*, 14, 2021.

[60] R. Li and X. Yang. De novo reconstruction of cell interaction landscapes from single-cell spatial transcriptome data with deeplinc. *Genome Biology*, 23(1), 2022.

[61] Z. Li, T. Wang, P. Liu, and Y. Huang. Spatialdm for rapid identification of spatially co-expressed ligand–receptor and revealing cell–cell communication patterns. *Nat. Commun.*, 14, 2023.

[62] M. Liao, Y. Liu, J. Yuan, Y. Wen, G. Xu, J. Zhao, et al. Single-cell landscape of bronchoalveolar immune cells in patients with covid-19. *Nature Medicine*, 26:842–844, 2020.

[63] A. Louvi, D. Accili, and A. Efstratiadis. Growth-promoting interaction of igf-ii with the insulin receptor during mouse embryonic development. *Dev. Biol.*, 189, 1997.

[64] M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Müller, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19:41–50, 2022.

[65] F. Ma, S. Zhang, L. Song, B. Wang, L. Wei, and F. Zhang. Applications and analytical tools of cell communication based on ligand-receptor interactions at single cell level. *Cell and Bioscience*, 11, 2021.

[66] J. C. Melms, J. Biermann, H. Huang, Y. Wang, A. Nair, S. Tagore, et al. A molecular single-cell lung atlas of lethal covid-19. *Nature*, 595:114–119, 2021.

[67] L. Michielsen, M. J. T. Reinders, and A. Mahfouz. Hierarchical progressive learning of cell identities in single-cell data. *Nat. Commun.*, 12, 2021.

[68] X. Mo, Q. Niu, A. Ivanov, Y. Tsang, C. Tang, C. Shu, Q. Li, K. Qian, A. Wahafu, S. Doyle, and et al. Systematic discovery of mutation-directed neo-protein-protein interactions in cancer. *Cell*, 185(11):1974–1985.e12, 2022.

[69] S. Mukherjee. Toll-like receptor 4 in covid-19: Friend or foe? *Future Virol.*, 17, 2022.

[70] M. Najafi. Effect of calprotectin subunit s100a9 on the expression and methylation of ocln in human melanoma cell line a-375. *Turkish J. Biol.*, 41, 2017.

[71] J. S. Nam. Effect of wnt3a on keratinocytes utilizing in vitro and bioinformatics analysis. *Int J. Mol. Sci.*, 15, 2014.

[72] F. Noël, L. Massenet-Regad, I. Carmi-Levy, A. Cappuccio, M. Grandclaudon, C. Trichot, Y. Kieffer, F. Mechta-Grigoriou, and V. Soumelis. Dissection of intercellular communication using the transcriptome-based framework icellnet. *Nature communications*, 12(1):1–16, 2021.

[73] E. Z. Ong, Y. F. F. Z. Chan, W. Y. Leong, N. M. Y. Lee, S. Kalimuddin, S. M. H. Mohideen, et al. A dynamic immune response shapes covid-19 progression. *Cell Host & Microbe*, 27:879–882.e2, 2020.

[74] L. Przybyla and J. Voldman. Probing embryonic stem cell autocrine and paracrine signaling using microfluidics. *Annual Review of Analytical Chemistry*, 5:293–315, 2012.

[75] S. Puram, I. Tirosh, A. Parikh, A. Patel, K. Yizhak, S. Gillespie, C. Rodman, and et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171(7):1611–1624.e24, 2017.

[76] M. Raredon, J. Yang, J. Garritano, M. Wang, D. Kushnir, J. Schupp, T. Adams, A. Greaney, K. Leiby, N. Kaminski, Y. Kluger, A. Levchenko, and L. Niklason. Computation and visualization of cell–cell signaling topologies in single-cell systems data using connectome. *Scientific Reports*, 12(1), 2022.

[77] X. Ren, W. Wen, X. Fan, W. Hou, B. Su, P. Cai, J. Li, Y. Liu, F. Tang, F. Zhang, et al. Covid-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell*, 184(7):1895–1913, 2021.

[78] X. Ren, G. Zhong, Q. Zhang, L. Zhang, Y. Sun, and Z. Zhang. Reconstruction of cell spatial organization from single-cell rna sequencing data based on ligand-receptor mediated self-assembly. *Cell Research*, 30(9):763–778, 2020.

[79] A. Ridiandries, J. T. M. Tan, and C. A. Bursill. The role of chemokines in wound healing. *Int J. Mol. Sci.*, 19, 2018.

[80] L. Ring, P. Neth, C. Weber, S. Steffens, and A. Faussner. b-catenin-dependent pathway activation by both promiscuous 'canonical' wnt3a-, and specific 'noncanonical' wnt4- and wnt5a-fzd receptor combinations with strong differences in lrp5 and lrp6 dependency. *Cell Signal*, 26, 2014.

[81] K. Räsänen and M. Herlyn. Paracrine signaling between carcinoma cells and mesenchymal stem cells generates cancer stem cell niche via epithelial-mesenchymal transition. *Cancer Discovery*, 2(9):775–777, 2012.

[82] M. Saitou. Complex phenotype of mice lacking occludin, a component of tight junction strands. *Mol. Biol. Cell*, 11, 2000.

[83] L. Sauteur. Cdh5/ve-cadherin promotes endothelial cell interface elongation via cortical actin polymerization during angiogenic sprouting. *Cell Rep.*, 9, 2014.

[84] M. T. Schaub, A. R. Benson, P. Horn, G. Lippner, and A. Jadbabaie. Multiscale dynamical embeddings of complex networks. *Physical Review E*, 99:062308, 2019.

[85] R. Sever and J. Brugge. Signal transduction in cancer. *Cold Spring Harbor Perspectives in Medicine*, 5(4), 2015.

[86] Y. Sha, Y. Qiu, P. Zhou, and Q. Nie. Reconstructing growth and dynamic trajectories from single-cell transcriptomics data. *Nat. Mach. Intell.*, 6, 2024.

[87] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[88] E. Stephenson, G. Reynolds, R. A. Botting, F. J. Calero-Nieto, M. D. Morgan, Z. K. Tuong, et al. Single-cell multi-omics analysis of the immune response in covid-19. *Nature Medicine*, 27:904–916, 2021.

[89] D. A. SuVerle, S. Yotsukura, S. Nomura, H. Aburatani, and K. Tsuda. Celltree: An r/bioconductor package to infer the hierarchical structure of cell populations from single-cell rna-seq data. *BMC Bioinforma.*, 17, 2016.

[90] V. Svensson, R. Vento-Tormo, and S. Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature Protocols*, 13(4):599–604, 2018.

[91] J. Tanevski, R. Flores, A. Gabor, D. Schapiro, and J. Saez-Rodriguez. Explainable multiview framework for dissecting spatial relationships from highly multiplexed data. *Genome Biology*, 23(1), 2022.

[92] R. Tang, C. Murray, I. Linde, N. Kramer, Z. Lyu, M. Tsai, L. Chen, and et al. A versatile system to record cell-cell interactions. *eLife*, 9:1, 2020.

[93] Z. Tang, T. Zhang, B. Yang, J. Su, and Q. Song. spaci: deciphering spatial cellular communications through adaptive graph model. *Brief. Bioinform*, 24, 2023.

[94] F. Tepekoy, G. Akkoyunlu, and R. Demir. The role of wnt signaling members in the uterus and embryo during pre-implantation and implantation. *Journal of Assisted Reproduction and Genetics*, 32(3):337–346, 2015.

[95] S. R. Tyler. Pyminer finds gene and autocrine-paracrine networks from human islet scrna-seq. *Cell Rep.*, 26, 2019.

[96] S. Umar. Ccl25 and ccr9 is a unique pathway that potentiates pannus formation by remodeling ra macrophages into mature osteoclasts. *Eur. J. Immunol.*, 51, 2021.

[97] A. Valdivia, A. M. Avalos, and L. Leyton. Thy-1 (cd90)-regulated cell adhesion and migration of mesenchymal cells: insights into adhesomes, mechanical forces, and signaling pathways. *Front Cell Dev. Biol.*, 11, 2023.

[98] B. Walker, Z. Cang, H. Ren, E. Bourgain-Chang, and Q. Nie. Deciphering tissue structure and function using spatial transcriptomics. *Communications Biology*, 5(1):220, 2022.

[99] B. L. Walker and Q. Nie. Nest: nested hierarchical structure identification in spatial transcriptomic data. *Nat. Commun.*, 14, 2023.

[100] B. Wang, J. Han, J. H. Elisseeff, and M. Demaria. The senescence-associated secretory phenotype and its physiological and pathological implications. *Nature Reviews Molecular Cell Biology*, pages 1–21, 2024.

[101] M. Wang et al. High-resolution 3d spatiotemporal transcriptomic maps of developing drosophila embryos and larvae. *Developmental Cell*, 57(10):1271–1283.e4, 2022.

[102] S. Wang, A. L. MacLean, and Q. Nie. Soptsc: similarity matrix optimization for clustering, lineage, and signaling inference. *bioRxiv*, page 168922, 2018.

[103] S. Wang, M. Zhao, F. Shao, W. Zhou, P. Xia, and P. Yang. Communication pattern changes along with declined igf1 of immune cells in covid-19 patients during disease progression. *Frontiers in immunology*, page 5862, 2022.

[104] X. Wang, A. Almet, and Q. Nie. Analyzing network diversity of cell–cell interactions in covid-19 using single-cell transcriptomics. *Frontiers in Genetics*, 13, 2022.

[105] X. Wang, A. A. Almet, and Q. Nie. The promising application of cell-cell interaction analysis in cancer from single-cell and spatial transcriptomics. *Semin Cancer Biol.*, 95, 2023.

[106] Y. Wang et al. italk: an r package to characterize and illustrate intercellular communication. *bioRxiv*, page 507871, 2019.

[107] S. L. Wolock, R. Lopez, and A. M. Klein. Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Systems*, 8(4):281–291, 2019.

[108] X. Xiong, H. Kuang, S. Ansari, T. Liu, J. Gong, S. Wang, X.-Y. Zhao, Y. Ji, C. Li, L. Guo, et al. Landscape of intercellular crosstalk in healthy and nash liver revealed by single-cell secretome gene analysis. *Molecular cell*, 75(3):644–660, 2019.

[109] A. Yu, Y. Li, I. Li, M. Ozawa, C. Yeh, A. Chiou, W. Trope, J. Taylor, J. Shrager, and S. Plevritis. Reconstructing codependent cellular cross-talk in lung adenocarcinoma using remi. *Science Advances*, 8, 2022.

[110] Y. Yuan and Z. Bar-Joseph. Gcng: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biology*, 21(1), 2020.

[111] J.-Y. Zhang, X.-M. Wang, X. Xing, Z. Xu, C. Zhang, J.-W. Song, et al. Single-cell landscape of immunological responses in patients with covid-19. *Nature Immunology*, 21:1107–1118, 2020.

[112] Y. Zhang, T. Liu, X. Hu, M. Wang, J. Wang, B. Zou, P. Tan, T. Cui, Y. Dou, L. Ning, Y. Huang, S. Rao, D. Wang, and X. Zhao. Cellcall: Integrating paired ligand-receptor and transcription factor activities for cell-cell communication. *Nucleic Acids Research*, 49(15):8520–8534, 2021.

[113] W. Zhao, K. G. Johnston, H. Ren, X. Xu, and Q. Nie. Inferring neuron-neuron communications from single-cell transcriptomics through neuronchat. *Nat. Commun.*, 14, 2023.

Appendix A

Additional file for Chapter 2

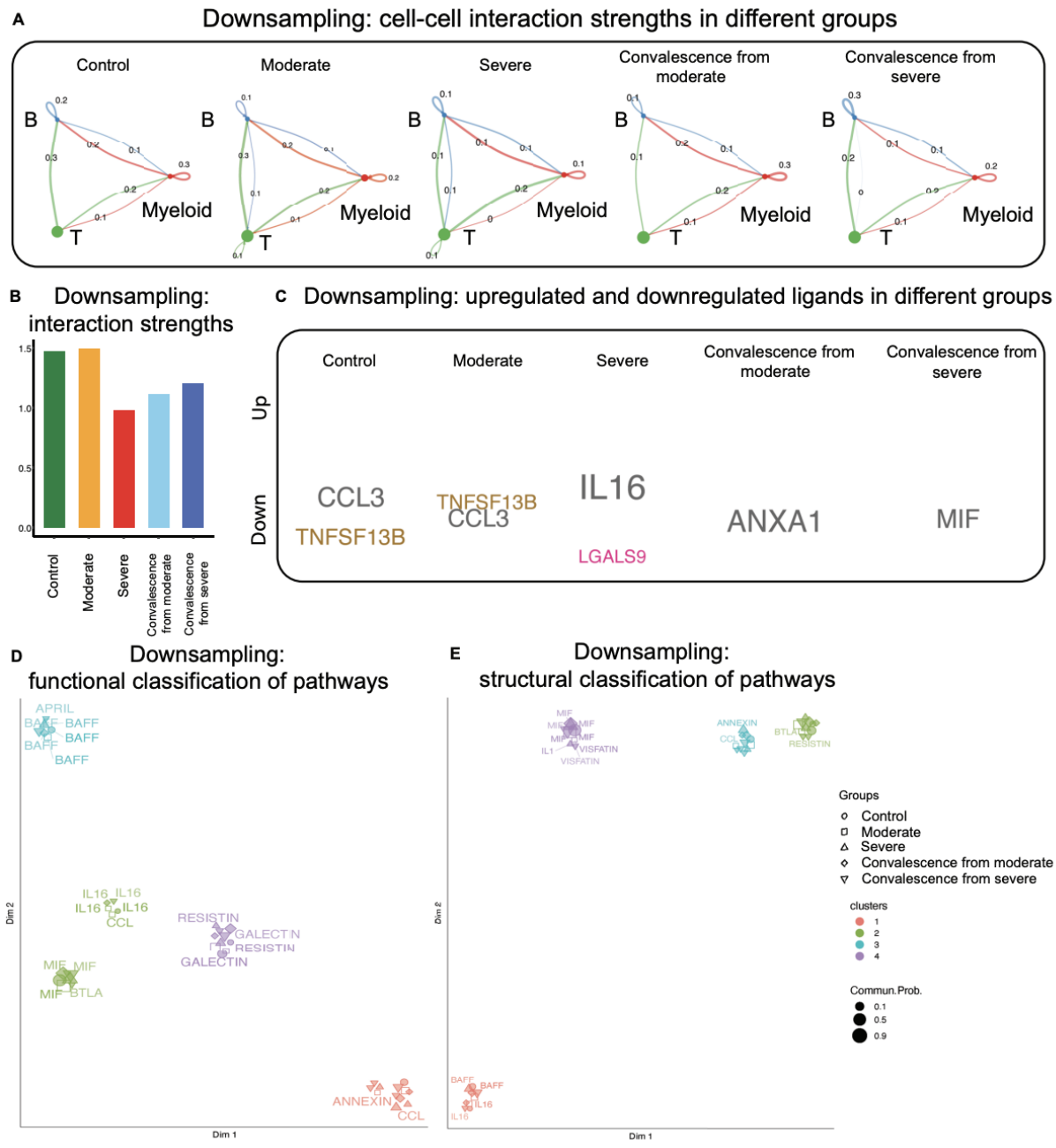


Figure A.1: **CellChat analysis of all pathways after downsampling.** (a). CCI interaction strengths of each downsampled condition, as obtained by random sampling of 80edge represents an interaction and its colour corresponds to the color of the sender cell type. (b). Total interaction strength in different downsampled groups. The total interaction strengths from the downsampled Control and Moderate groups are higher. (c). Upregulated and downregulated ligands for each downsampled condition group (d). Projection of signaling pathways onto a two-dimensional manifold according to their functional similarity. Each dot represents the interaction network of one signaling pathway. Different colors represent different inferred groups of signaling pathways. (e). Projection of signaling pathways onto a twodimensional manifold according to their structural similarity. Each dot represents the interaction network of one signaling pathway. Different colors represent different inferred groups of signaling pathways.

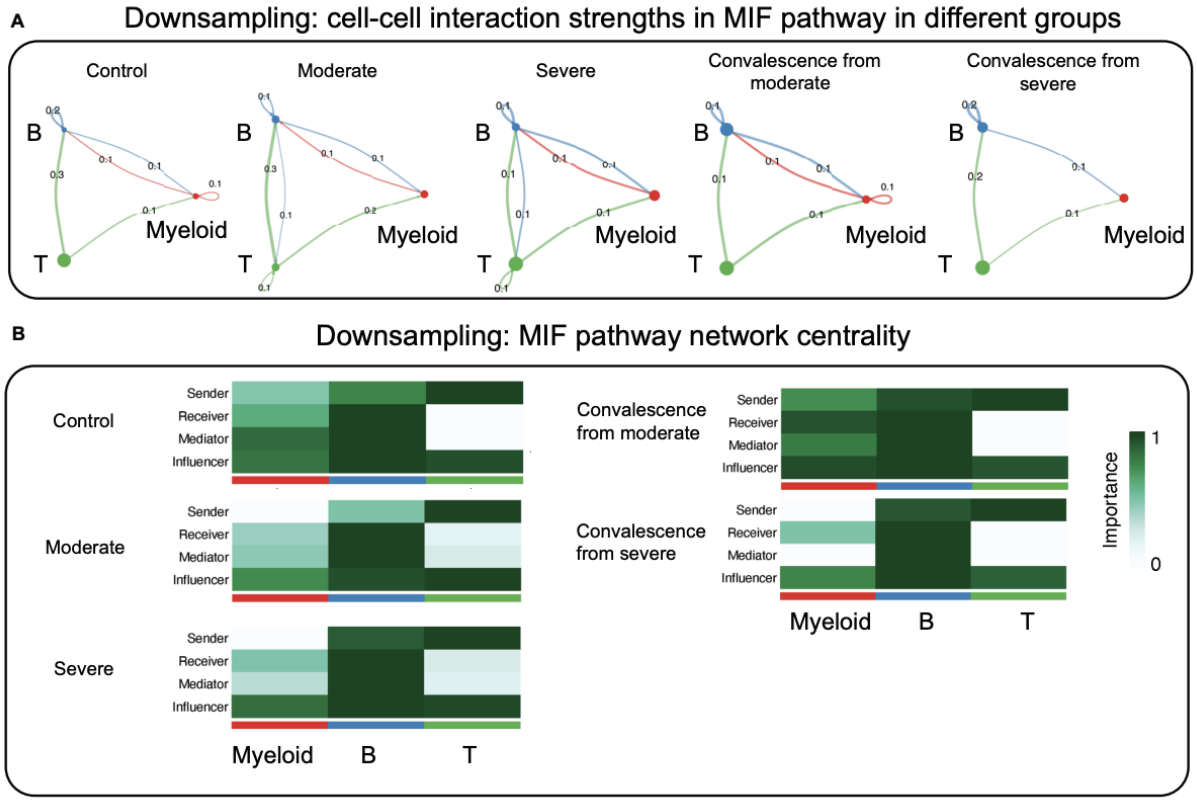


Figure A.2: **CellChat analysis of MIF pathway after downsampling.** (a). CCI strengths of each downsampled condition group. Each edge represents an interaction, where the color of the edge matches the sender cell type. (b). Network centrality scores of downsampled condition groups in MIF pathway. In downsampled severe and convalescence from severe groups, Myeloid cells are not implied to be mediators.

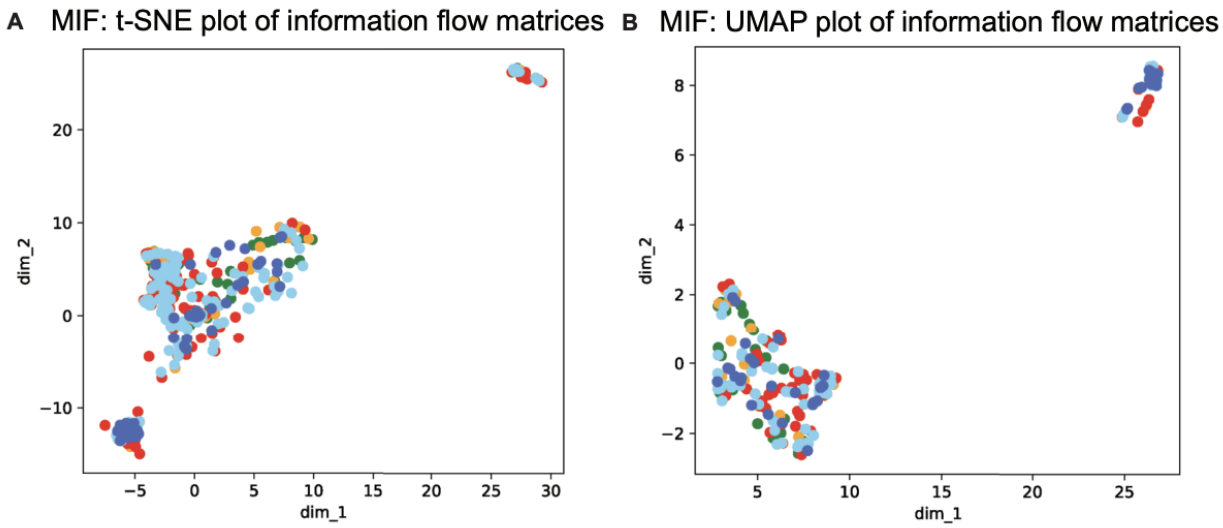


Figure A.3: **Visualization of information flow using nonlinear embeddings.** (a). t-SNE plot of information flow matrices of three cell types (Myeloid, B, T) in MIF pathway. (b). UMAP plot of information flow matrices of three cell types (Myeloid, B, T) in MIF pathway.

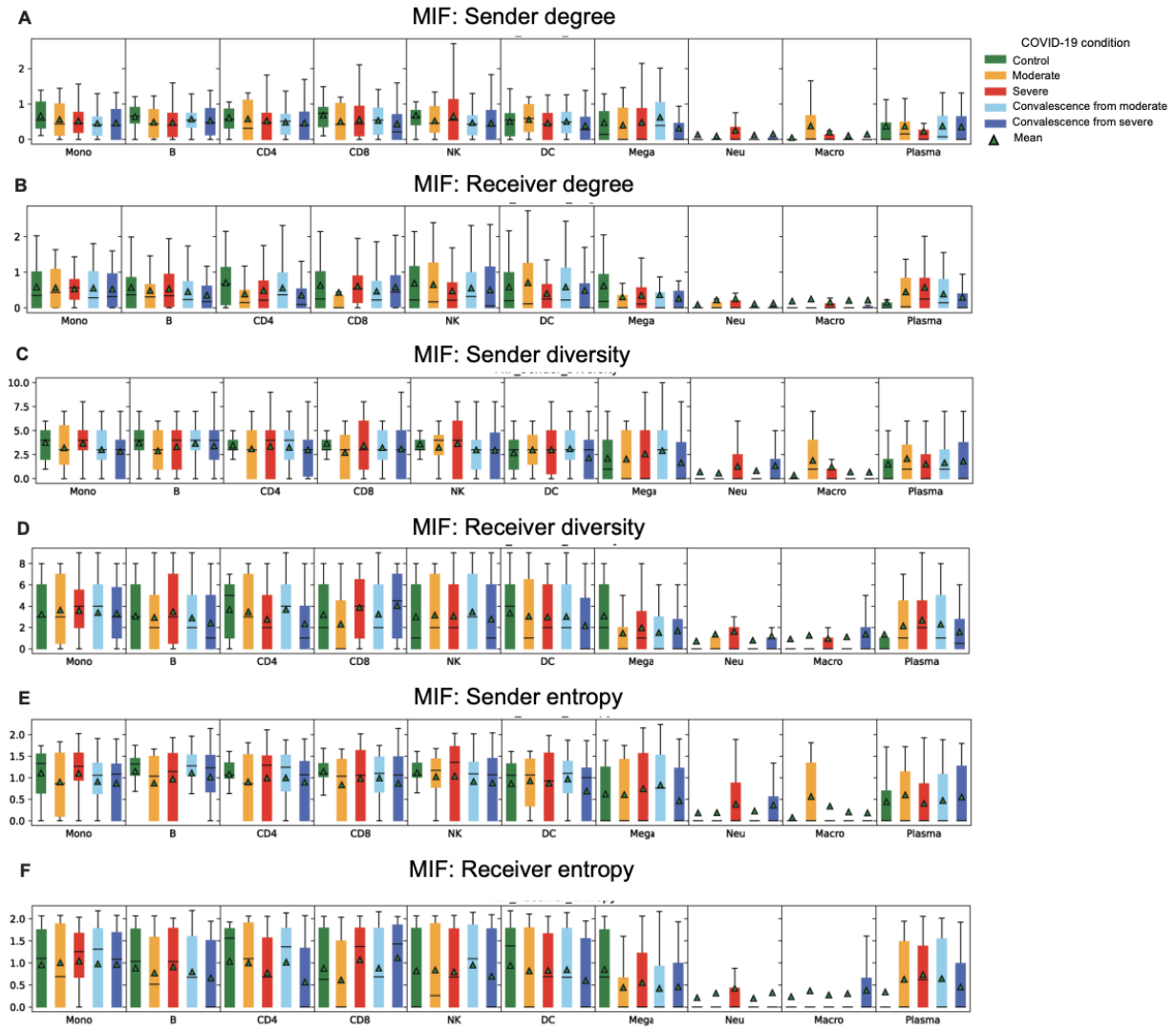


Figure A.4: **Distribution of ten cell types diversity statistics in different COVID-19 conditions are visualized.** (a). Distribution of sender degree in MIF pathway. (b). Distribution of receiver degree in MIF pathway. (c). Distribution of sender diversity in MIF pathway. (d). Distribution of receiver diversity in MIF pathway. The patterns are similar to receiver diversity. (e). Distribution of sender entropy in MIF pathway. F. Distribution of receiver entropy in MIF pathway.

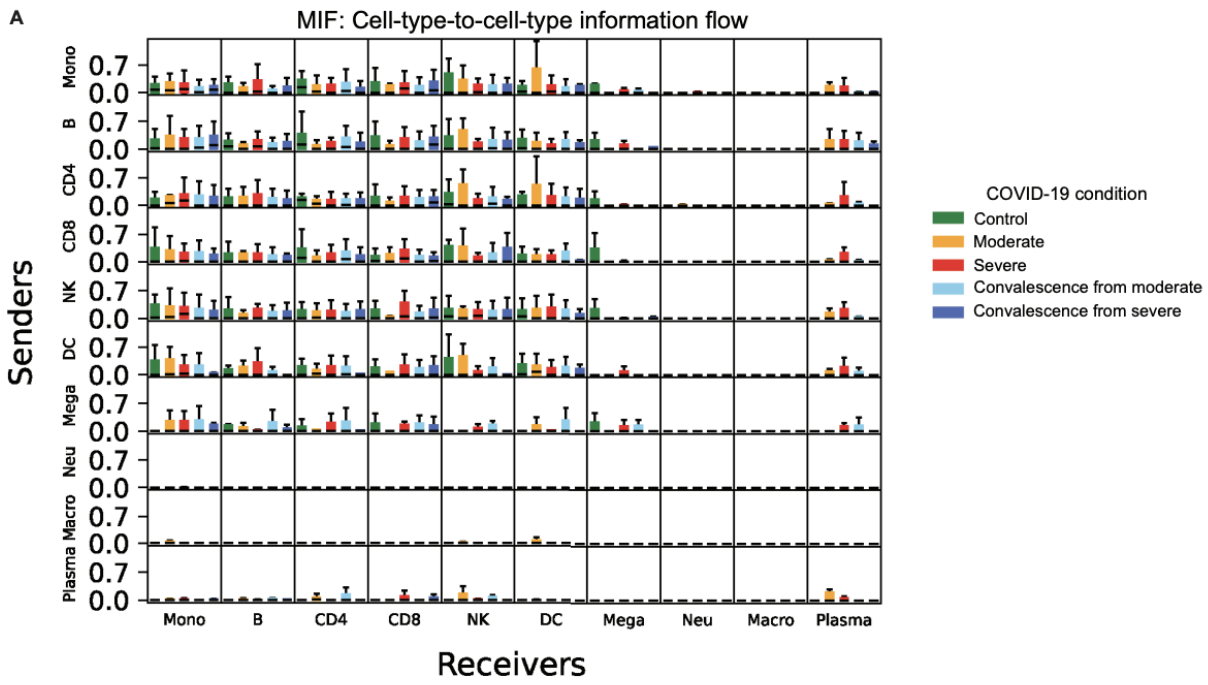


Figure A.5: **Distribution of node-to-node information flow statistic of ten cell types in MIF pathway in different COVID-19 conditions and comparison of non-linear projection of information flow proximity matrices of three cell types.** (a). Distribution of node-to-node information flow in MIF pathway in five condition groups.

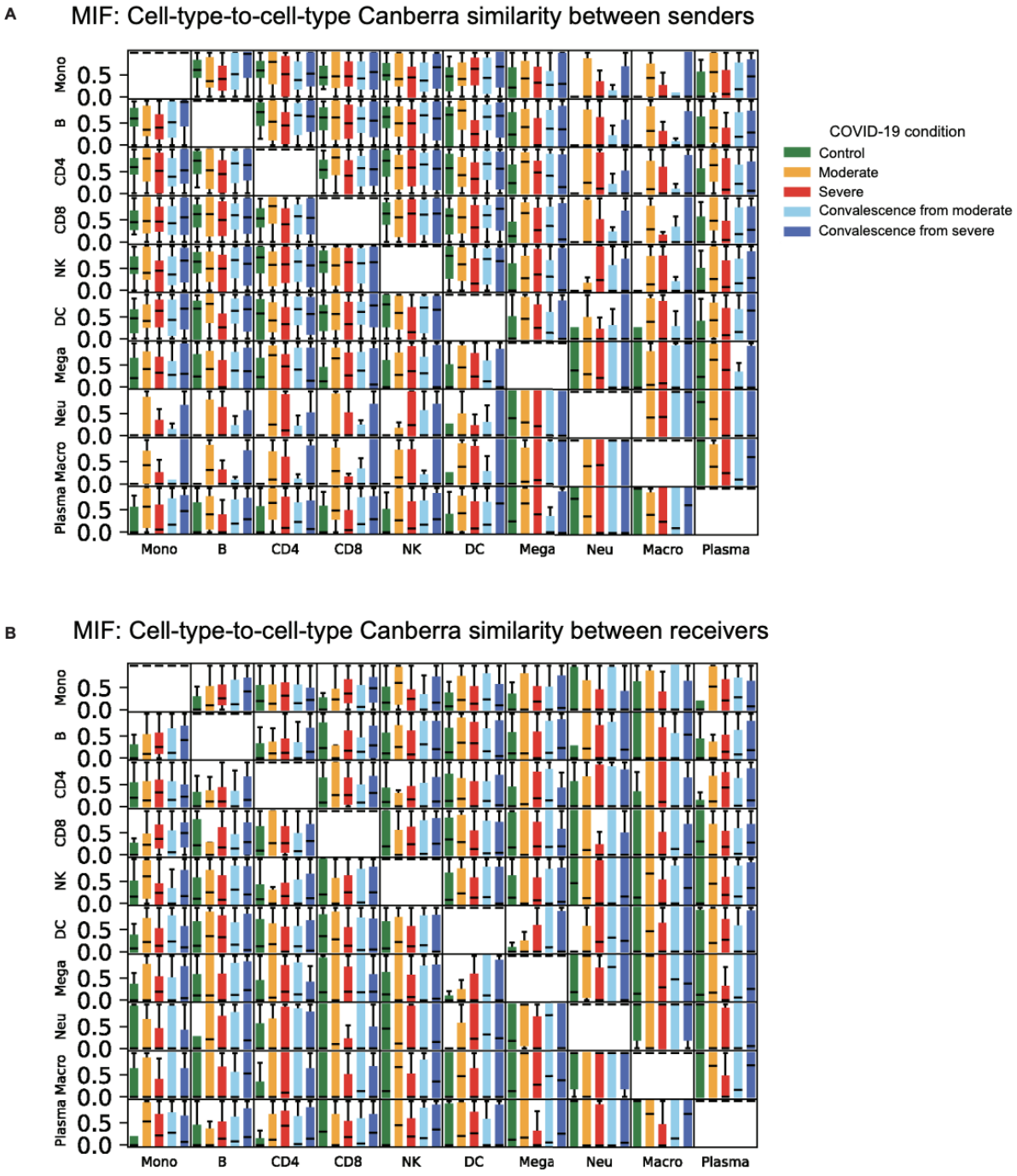


Figure A.6: **Distribution of node-to-node Canberra similarity of ten cell types in different COVID-19 conditions are visualized.** (a). Distribution of node-to-node Canberra similarity between senders in five condition groups. (b). Distribution of node-to-node Canberra similarity between receivers in five condition groups.

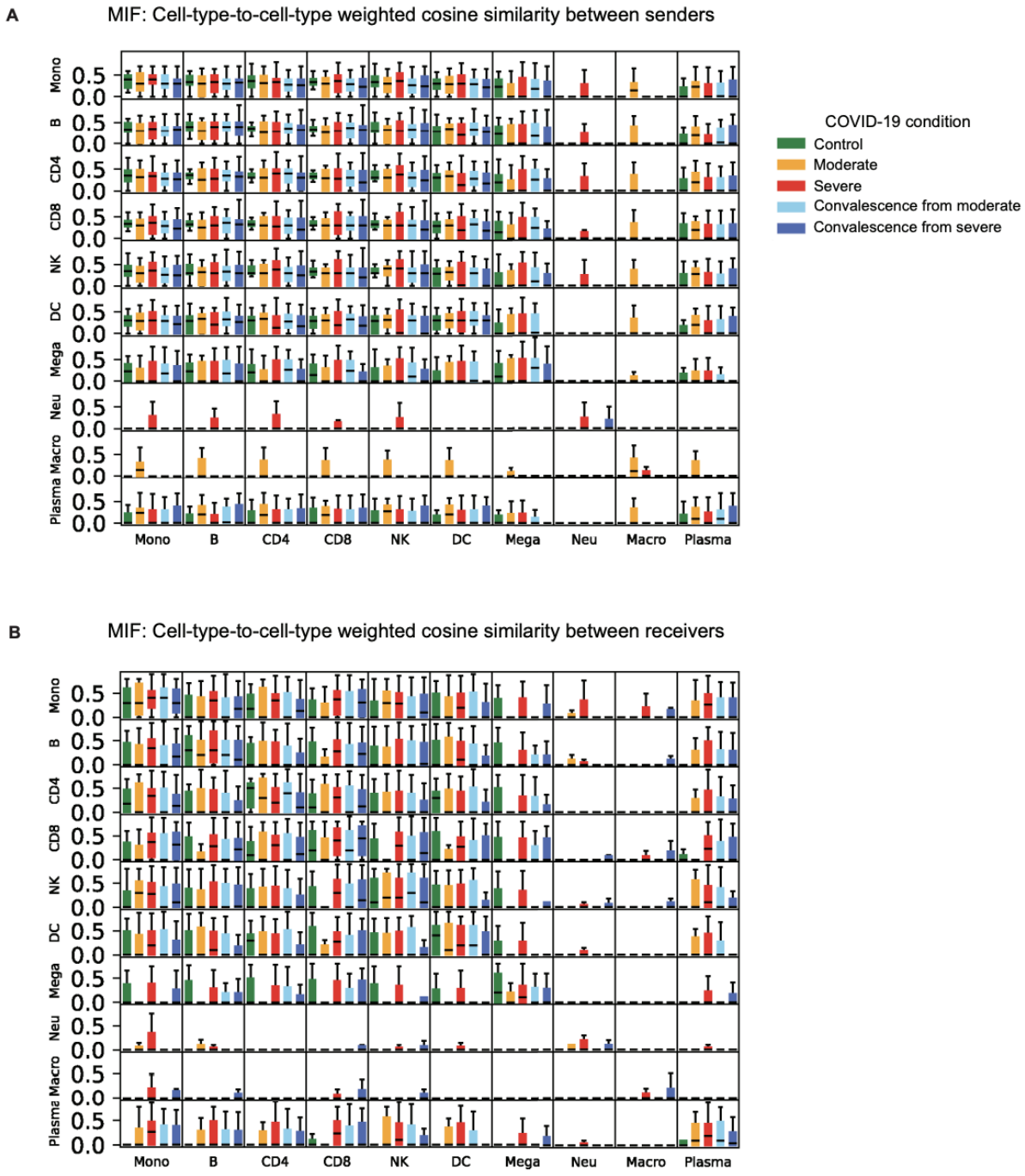


Figure A.7: **Distribution of node-to-node weighted cosine similarity of ten cell types in different COVID-19 conditions are visualized.** (a). Distribution of node-to-node weighted cosine similarity between senders in five condition groups. (b). Distribution of node-to-node weighted cosine similarity between receivers in five condition groups.

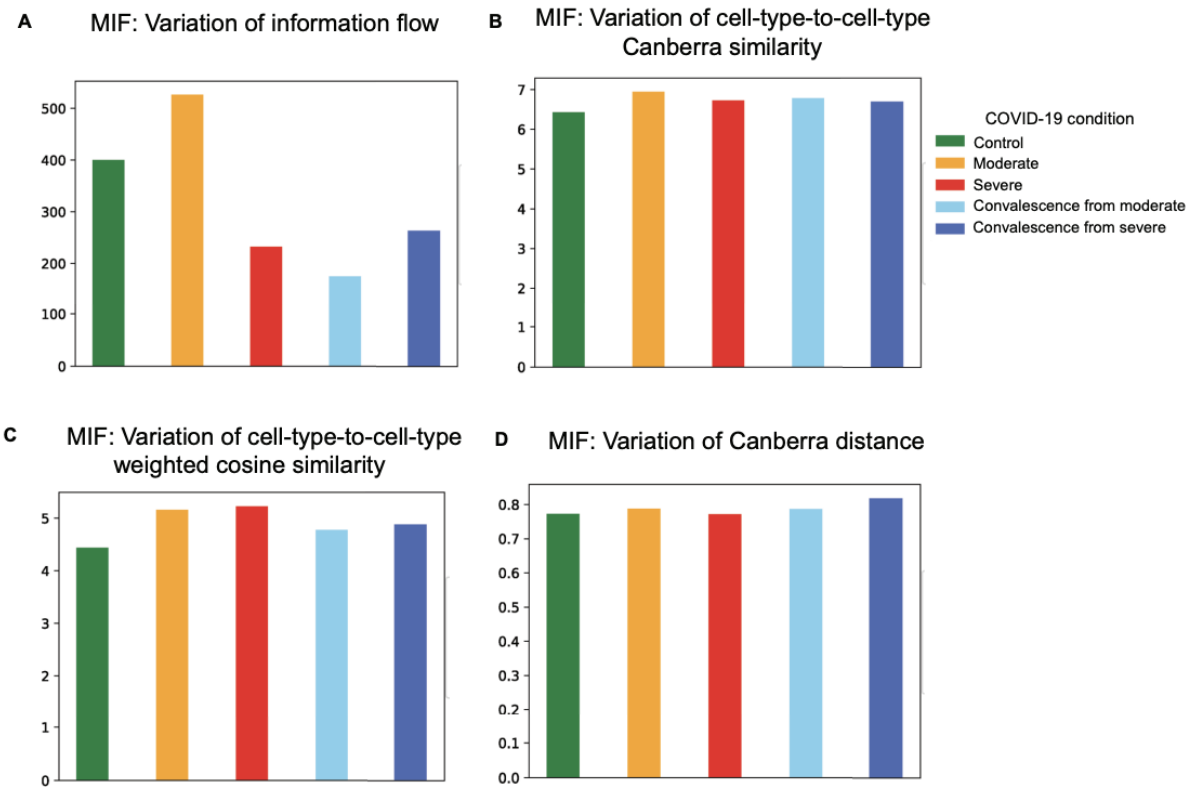


Figure A.8: **Network variability based on four distance measures in ten cell type networks.** (a). Network variability based on information flow distance. (b). Network variability based on cell-type-to-cell-type Canberra distance. (c). Network variability based on cell-type-to-cell-type weighted cosine similarity. (d). Network variability based on graph-to-graph Canberra distance.

Appendix B

Additional File for Chapter 3

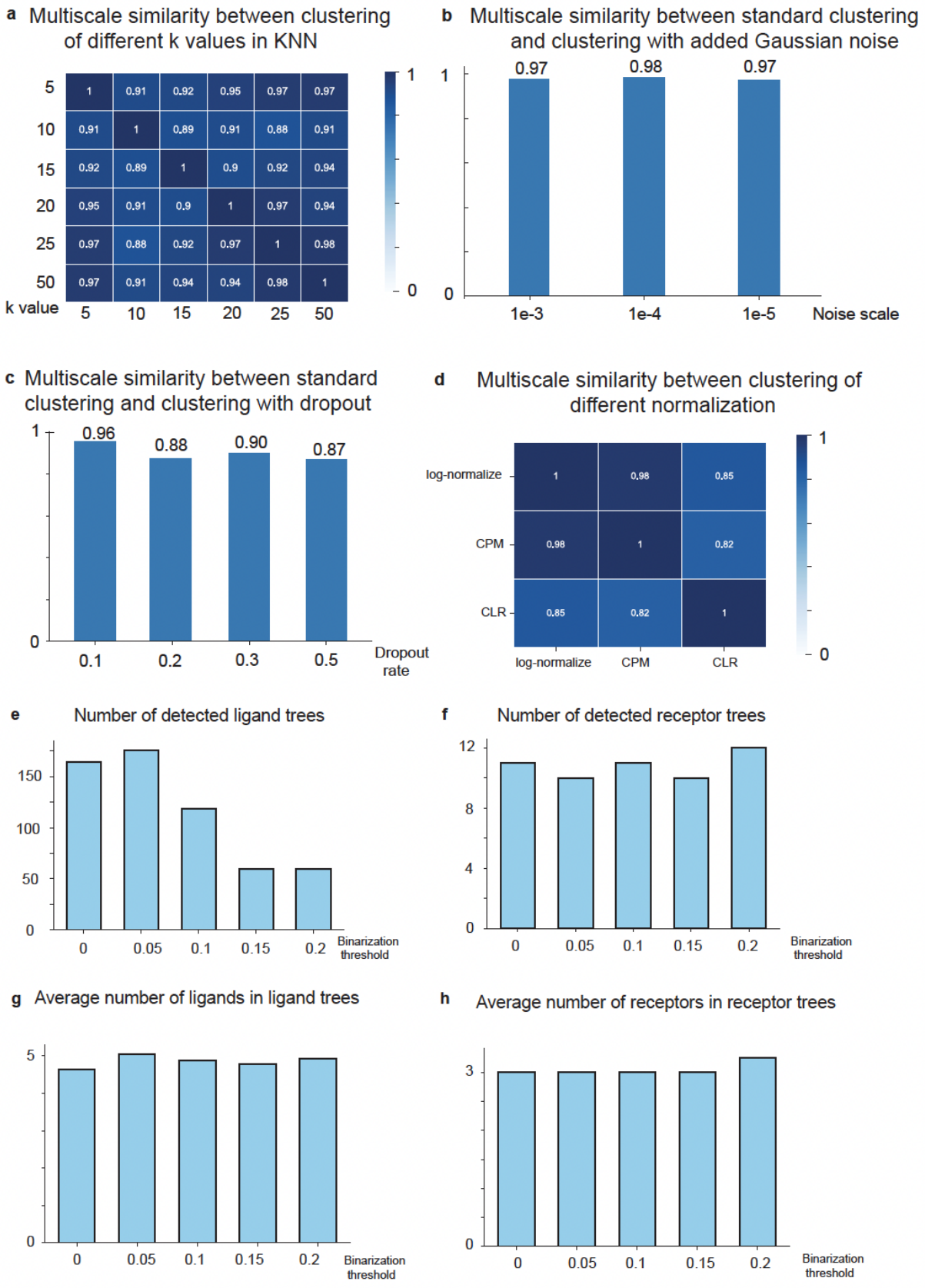


Figure B.1: **Robustness test of CrossChatH with respect to different parameters.** See next page for detailed caption.

Figure B.1 (continued). (a) Robustness test of CrossChatH with respect to selection of K in the K-nearest neighbor algorithm. (b) Robustness test of CrossChatH with respect to noise. (c) Robustness test of CrossChatH with respect to dropout. (d) Robustness test of CrossChatH with respect to different normalization. (e) Number of detected ligand trees with varying threshold. (f) Number of detected receptor trees with varying threshold. (g) Average number of ligands in ligand trees under varying threshold. (h) Average number of ligands in receptor trees under varying threshold.

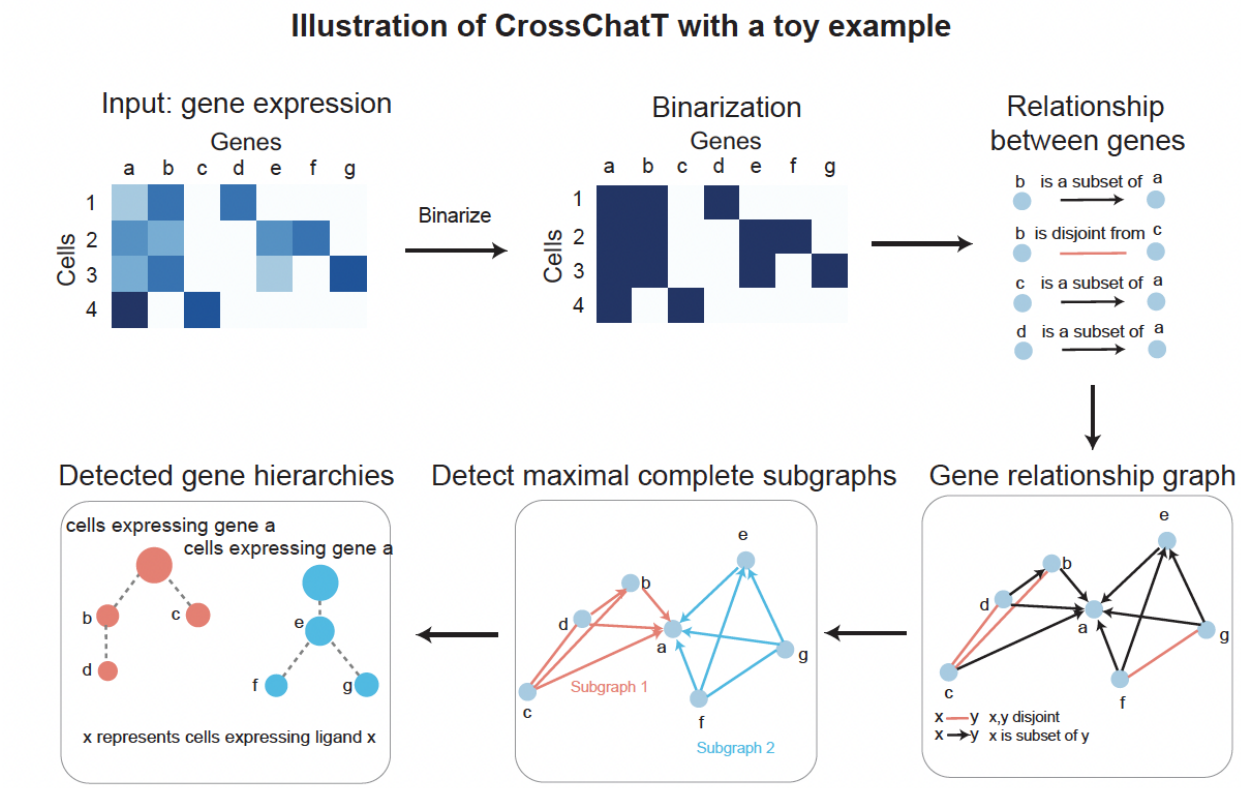
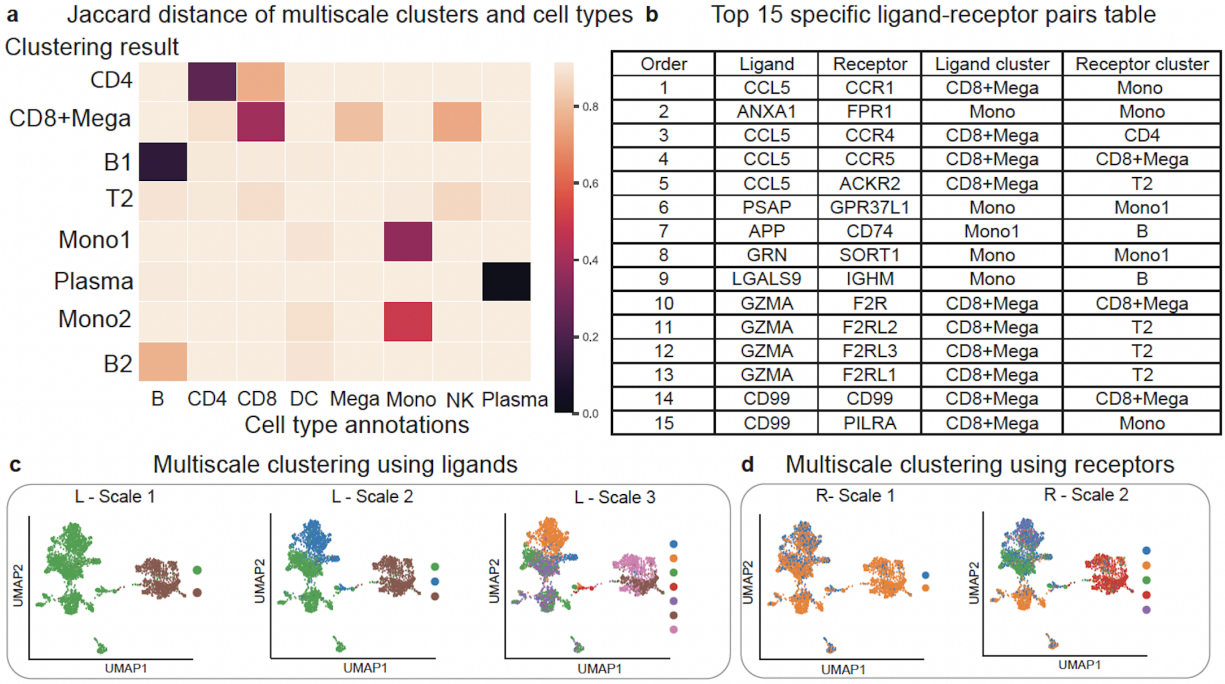


Figure B.2: **Detailed illustration of CrossChatT.** First, binarization on the gene expression matrix is performed in order to find the support for each ligand/receptor. Next, extract relationship between genes based on the binarization matrix, and build gene relationship graph. Based on the gene relationship graph, all maximal complete subgraphs are detected. Each detected subgraph forms a detected gene hierarchy.

Appendix C

Additional File for Chapter 4



Order	Ligand	Receptor	Ligand cluster	Receptor cluster
1	CCL5	CCR1	CD8+Mega	Mono
2	ANXA1	FPR1	Mono	Mono
3	CCL5	CCR4	CD8+Mega	CD4
4	CCL5	CCR5	CD8+Mega	CD8+Mega
5	CCL5	ACKR2	CD8+Mega	T2
6	PSAP	GPR37L1	Mono	Mono1
7	APP	CD74	Mono1	B
8	GRN	SORT1	Mono	Mono1
9	LGALS9	IGHM	Mono	B
10	GZMA	F2R	CD8+Mega	CD8+Mega
11	GZMA	F2RL2	CD8+Mega	T2
12	GZMA	F2RL3	CD8+Mega	T2
13	GZMA	F2RL1	CD8+Mega	T2
14	CD99	CD99	CD8+Mega	CD8+Mega
15	CD99	PILRA	CD8+Mega	Mono

Figure C.1: **Application of CrossChatH to PBMC cells sampled from COVID-19 patients.** (a) Jaccard distance between hierarchical clusters at finest scale and cell types. (b) Top 15 ligand-receptor pairs with highest specificity. Ligand cluster, receptor cluster correspond to their most specific ligand/receptor cluster. (c) Hierarchical clustering using only ligands as input gives three scales of clusters. (d) Hierarchical clustering using only receptors as input gives two scales of clusters.

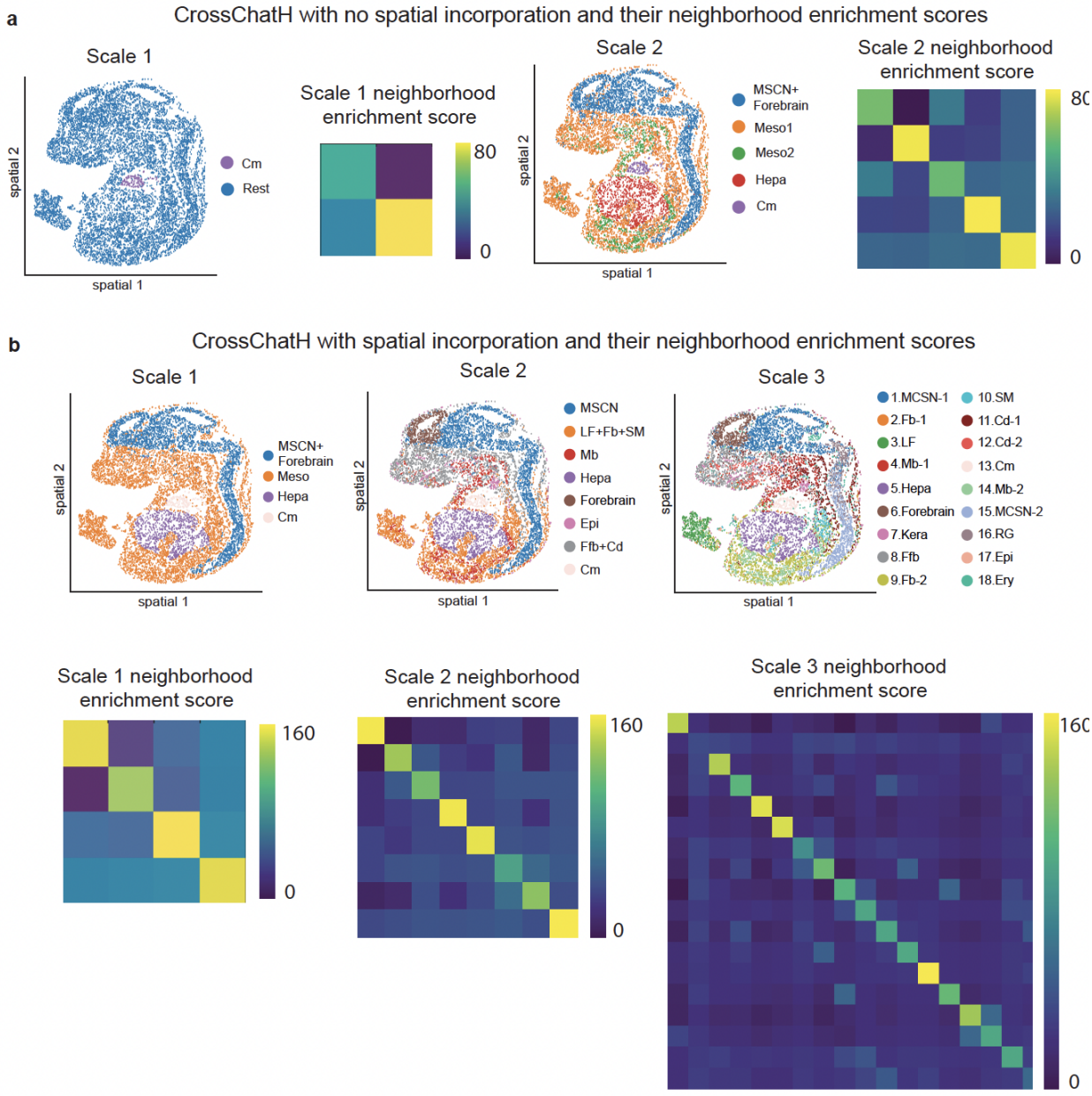
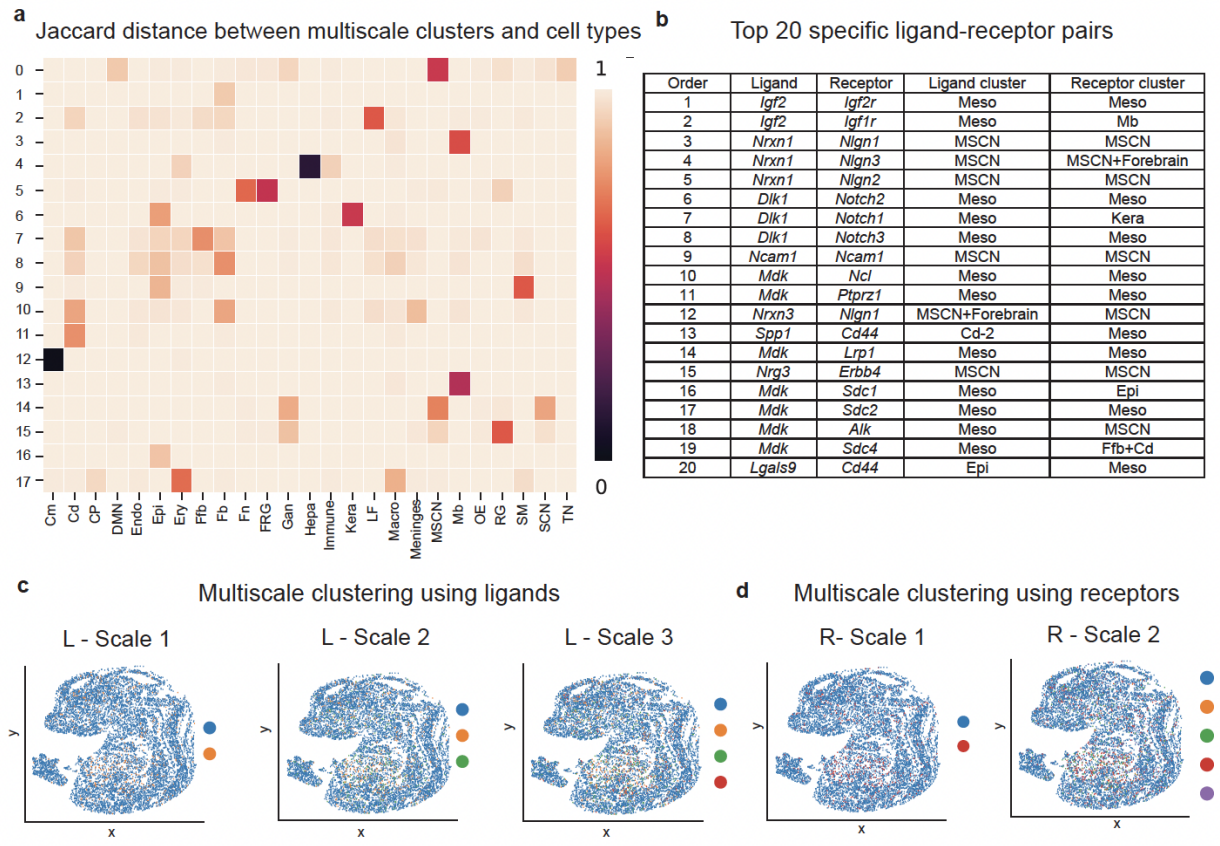


Figure C.2: **Demonstration of spatial incorporation for CrossChatH-S.** (a) CrossChatH with no spatial incorporation and their neighborhood enrichment scores (b) CrossChatH with spatial incorporation and their neighborhood enrichment scores. The detected clusters using CrossChatH-S which incorporates spatial information are more spatially coherent, compared to the clusters generated without spatial incorporation. This observation can be shown by higher neighborhood enrichment score of the spatial clusters detected by CrossChatH-S.



Order	Ligand	Receptor	Ligand cluster	Receptor cluster
1	*Igf2*	*Igf2r*	Meso	Meso
2	*Igf2*	*Igf1r*	Meso	Mb
3	*Nrxn1*	*Nlgn1*	MSCN	MSCN
4	*Nrxn1*	*Nlgn3*	MSCN	MSCN+Forebrain
5	*Nrxn1*	*Nlgn2*	MSCN	MSCN
6	*Dlk1*	*Notch2*	Meso	Meso
7	*Dlk1*	*Notch1*	Meso	Kera
8	*Dlk1*	*Notch3*	Meso	Meso
9	*Ncam1*	*Ncam1*	MSCN	MSCN
10	*Mdk*	*Ncl*	Meso	Meso
11	*Mdk*	*Ptprz1*	Meso	Meso
12	*Nrxn3*	*Nlgn1*	MSCN+Forebrain	MSCN
13	*Spp1*	*Cd44*	Cd-2	Meso
14	*Mdk*	*Lrp1*	Meso	Meso
15	*Nrg3*	*Erbb4*	MSCN	MSCN
16	*Mdk*	*Sdc1*	Meso	Epi
17	*Mdk*	*Sdc2*	Meso	Meso
18	*Mdk*	*Alk*	Meso	MSCN
19	*Mdk*	*Sdc4*	Meso	Ffb+Cd
20	*Lgals9*	*Cd44*	Epi	Meso

Figure C.3: **Application of CrossChatH-S to Stereo-seq mouse embryo at day E16.5 data.** (a) Jaccard distance between hierarchical clusters at finest scale and cell types of mouse embryo at day E16.5. (b) Top 20 ligand-receptor pairs with highest specificity. Ligand cluster, receptor cluster correspond to their most specific ligand/receptor cluster. (c) Hierarchical clustering using only ligands as input gives three scales of clusters. (d) Hierarchical clustering using only receptors as input gives two scales of clusters.

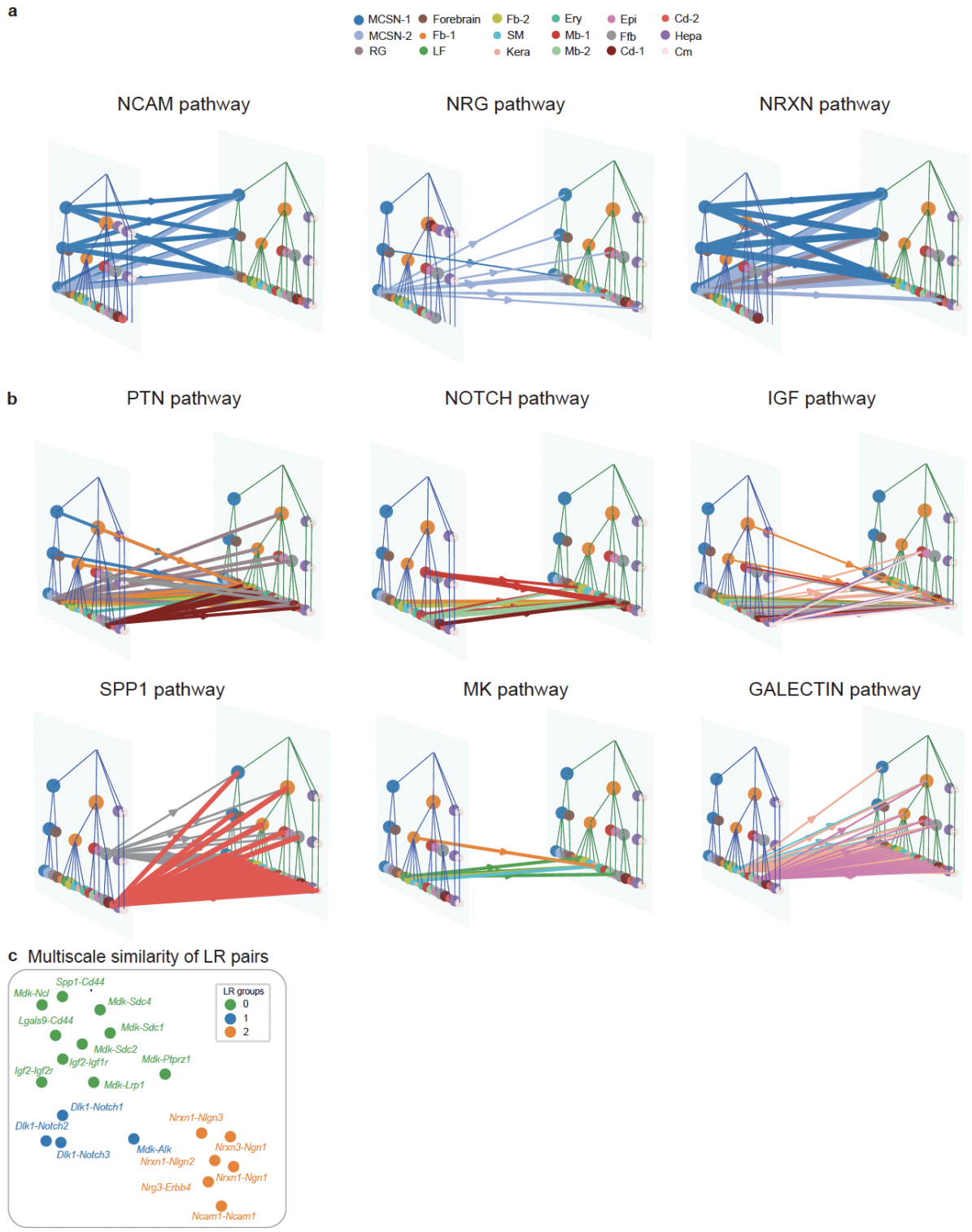


Figure C.4: **Cell-cell communications on Stereo-seq mouse embryo at E16.5 and multiscale similarity of ligand-receotir interactions.** See next page for detailed caption.

Figure C.4 (continued). (a) Cell-cell communications under Neural Cell Adhesion Molecule (NCAM), Neuregulin (NRG), Neurexin (NRXN) pathways. These pathways are associated with signaling mainly sent from brain and spinal cord neurons. (b) Cell-cell communications under Pleiotrophin (PTN), NOTCH, Insulin-like growth factor (IGF), Secreted phosphoprotein 1 (SPP1), Midkine (MK), GALECTIN pathways. These pathways are associated with signaling mainly sent from the rest of cells other than neurons. (c) Similarity of the top 20 most specific ligand-receptor interactions under hierarchical clusters. Group 1 includes interactions from MDK and IGF pathways, performing cell growth functions. Group 2 mainly contains interactions from Dlk1 to Notch receptors. Group 3 contains signaling within neurons.

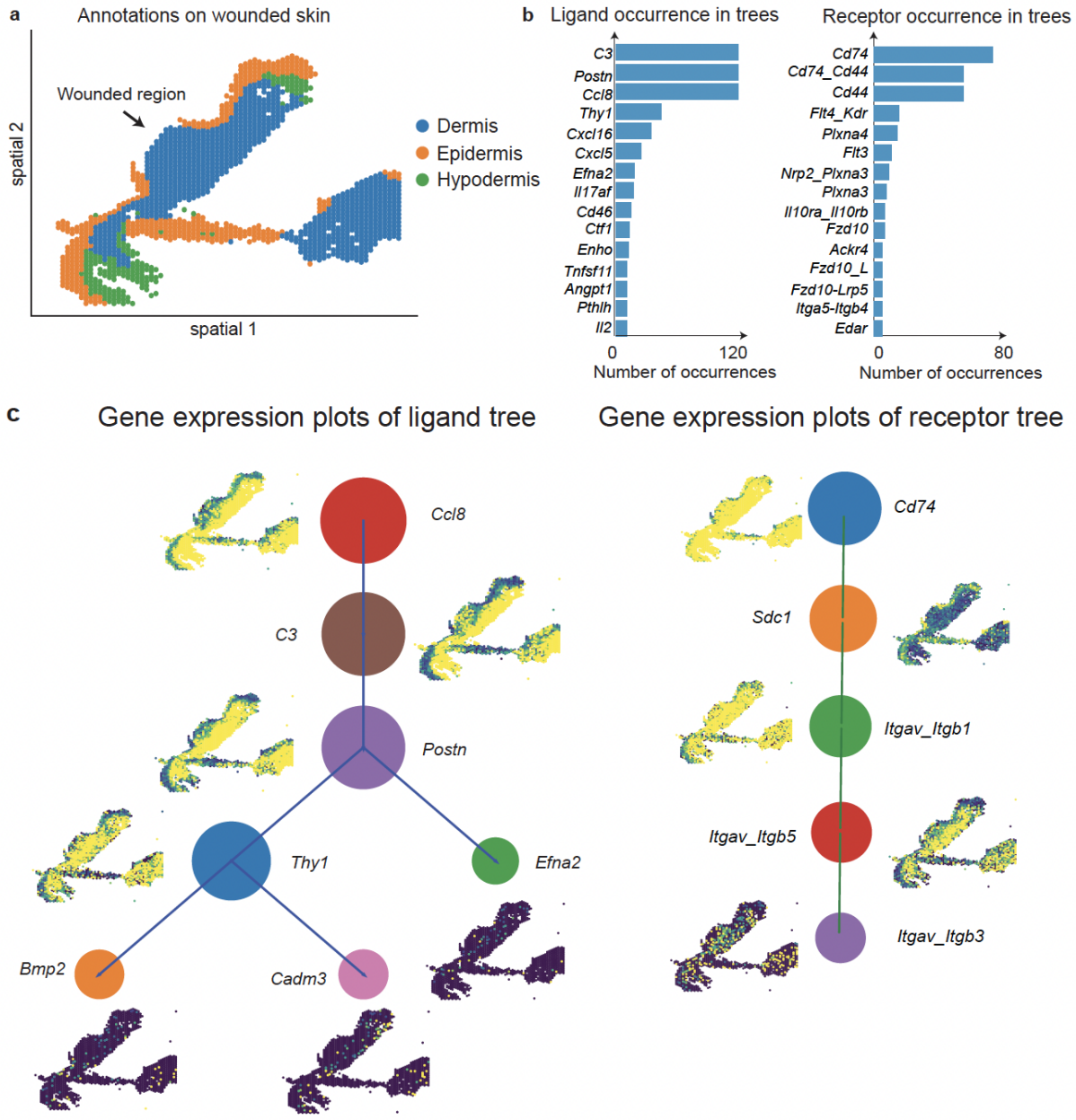


Figure C.5: **Cell-cell communication inferred from 10x Visium mouse skin.** See next page for detailed caption.

Figure C.5 (continued). (a) Annotations of cells in Visium mouse skin data at post-wound day 7. (b) Ligand/receptor occurrences in ligand/receptor trees detected using CrossChatT-S. (c) Gene expression of ligands and receptors in interaction tree 1 (see Fig. 6f). Cells expressing ligands/receptors at the top include cells containing ligands/receptors at the bottom. Cells expressing ligands/receptors at the same scale are disjoint from each other.

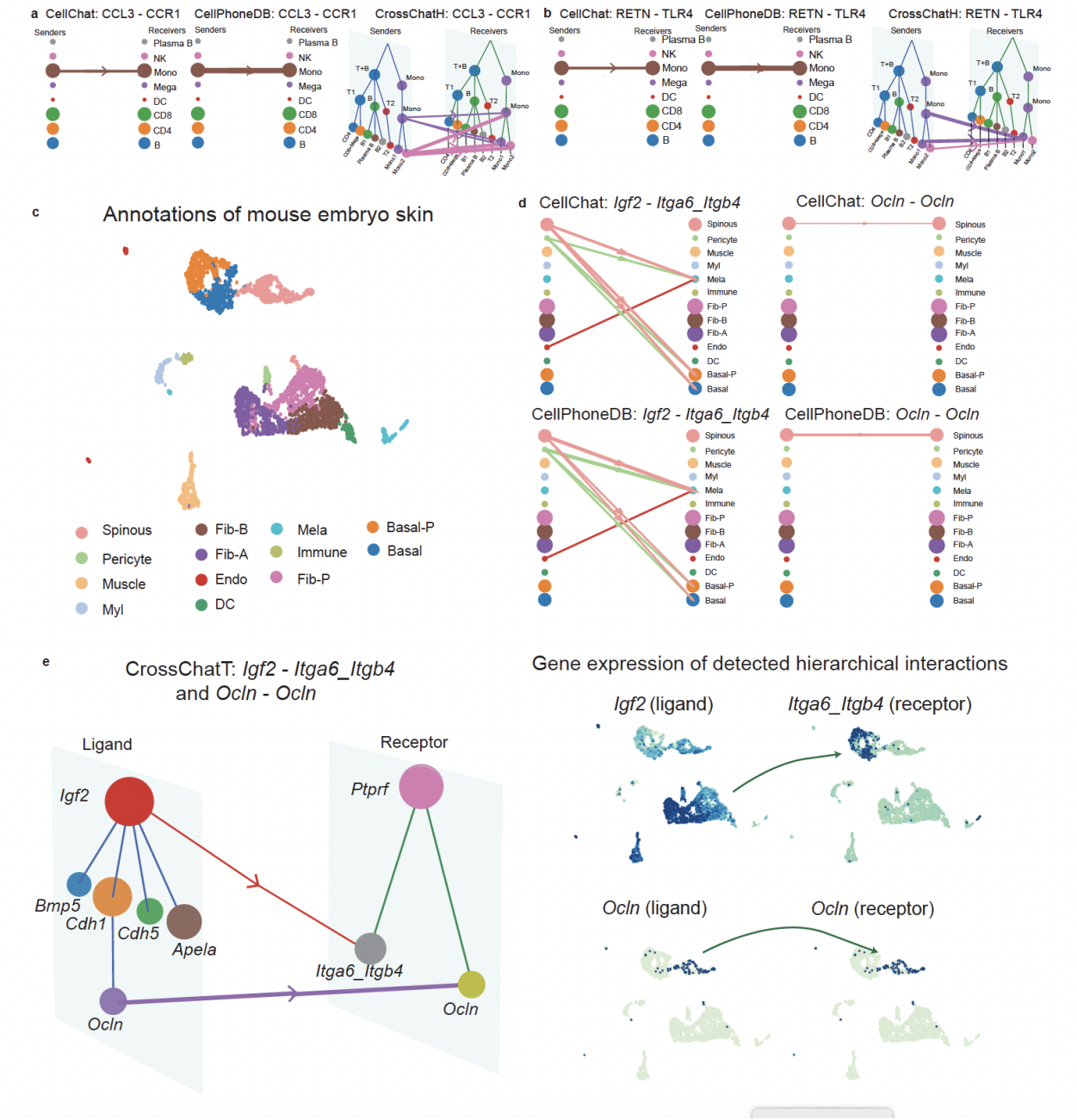


Figure C.6: **Comparison of CrossChat with existing CCC methods.** (a) CrossChat detects more precise signaling by assigning CCL3-CCR1 signaling to a subset of monocytes, Mono2. (b) CrossChat detects more precise signaling by assigning the receivers of RETN-TLR4 signaling to a subset of monocytes, Mono1. (c) Cell type annotations of a mouse embryo skin data. (d) Visualization of Igf2-Itga6_Itgb4, Ocln-Ocln interactions calculated using CellChat. e Visualization of Igf2-Itga6_Itgb4, Ocln-Ocln interactions calculated using CrossChatT.

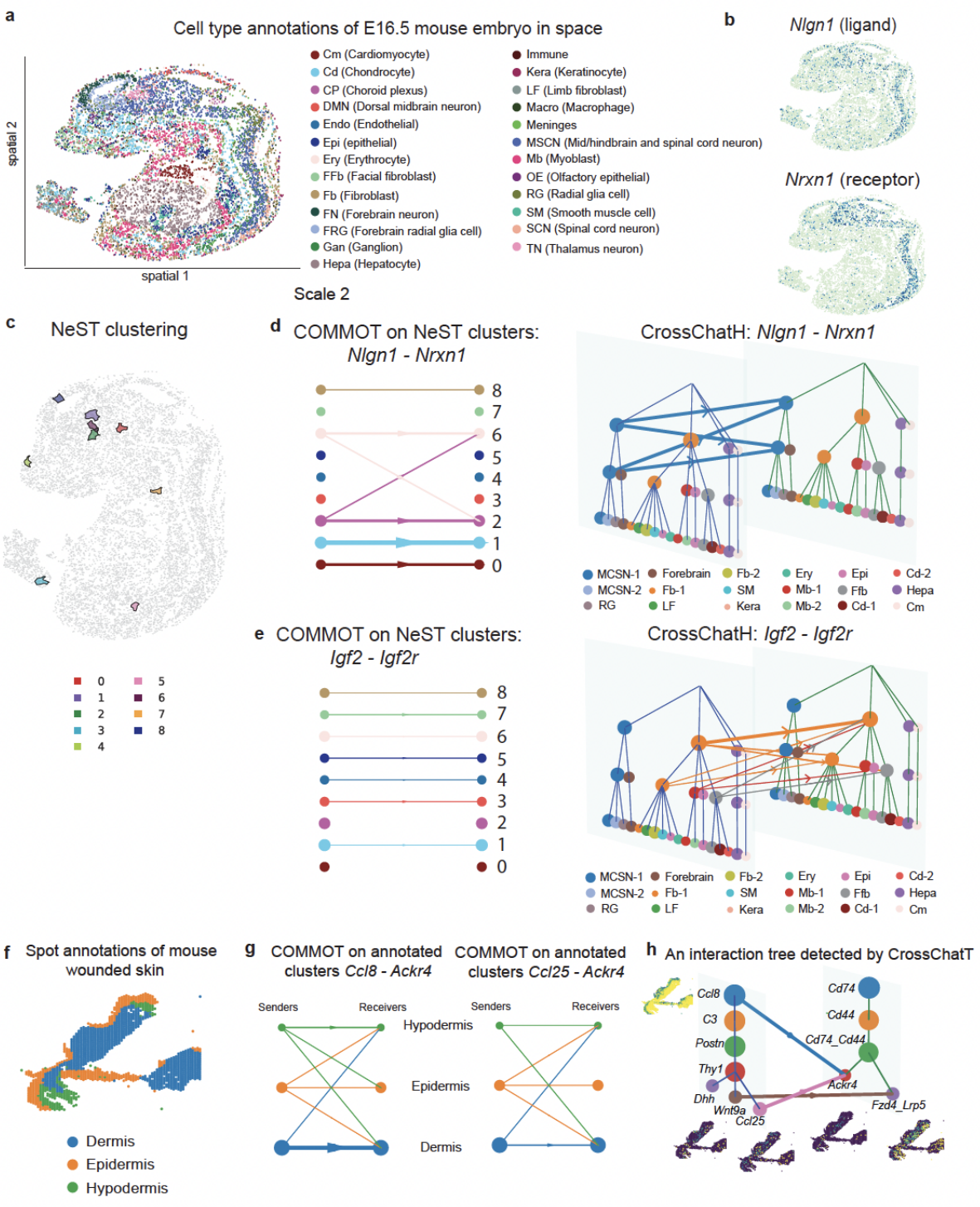


Figure C.7: **Comparison of CrossChat-S with existing spatial CCC methods.**

Figure C.7 (continued). (a) Cell type annotations of spatial mouse embryo dataset in space. (b) Gene expression of Nlgn1 ligand and Nrxn1 receptor in space. (c) Spatial clusters detected by NeST. (d) left: Visualization of Nlgn1-Nrxn1 interactions on NeST clusters calculated using COMMOT. right: Visualization of Nlgn1-Nrxn1 interactions on CrossChatH clusters calculated using COMMOT. (e) left: Visualization of Igf2-Igf2r interactions on NeST clusters calculated using COMMOT. right: Visualization of Igf-Igf2r interactions on CrossChatH clusters calculated using COMMOT. (f) Spot annotations of mouse wounded skin data in space. (g) Interaction of Ccl8–Ackr4, Ccl25–Ackr4 calculated using COMMOT. (h) An interaction tree detected by CrossChatT that includes Ccl8–Ackr4 and Ccl25–Ackr4 signaling.

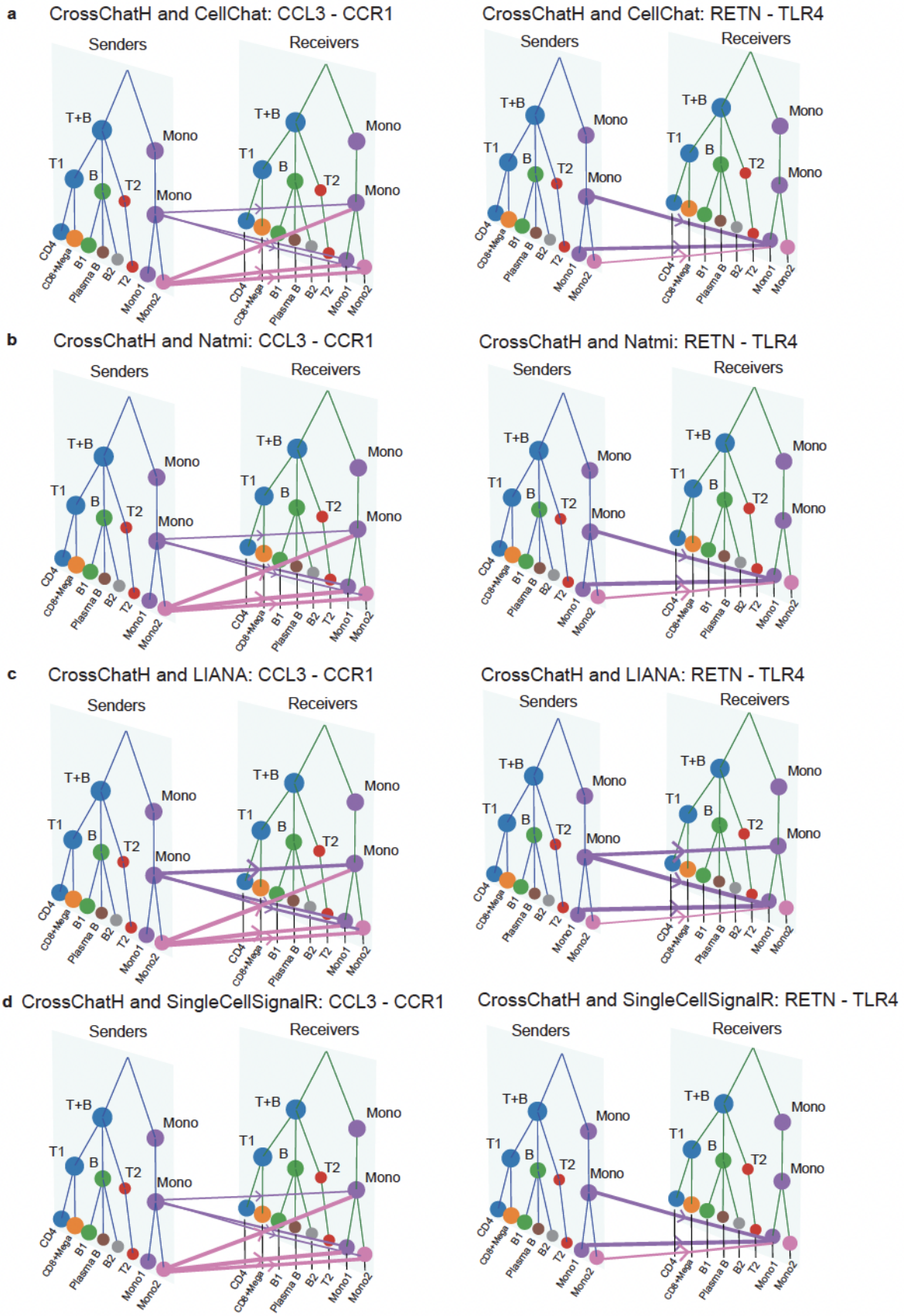


Figure C.8: **Comparison of CrossChatH analysis with different CCC packages.** See next page for detailed caption.

Figure C.8 (continued). (a) CCC detected by CrossChatH using CellChat (b) CCC detected by CrossChatH using NATMI (c) CCC detected by CrossChatH using LIANA (d) CCC detected by CrossChatH using SingleCellSignalR

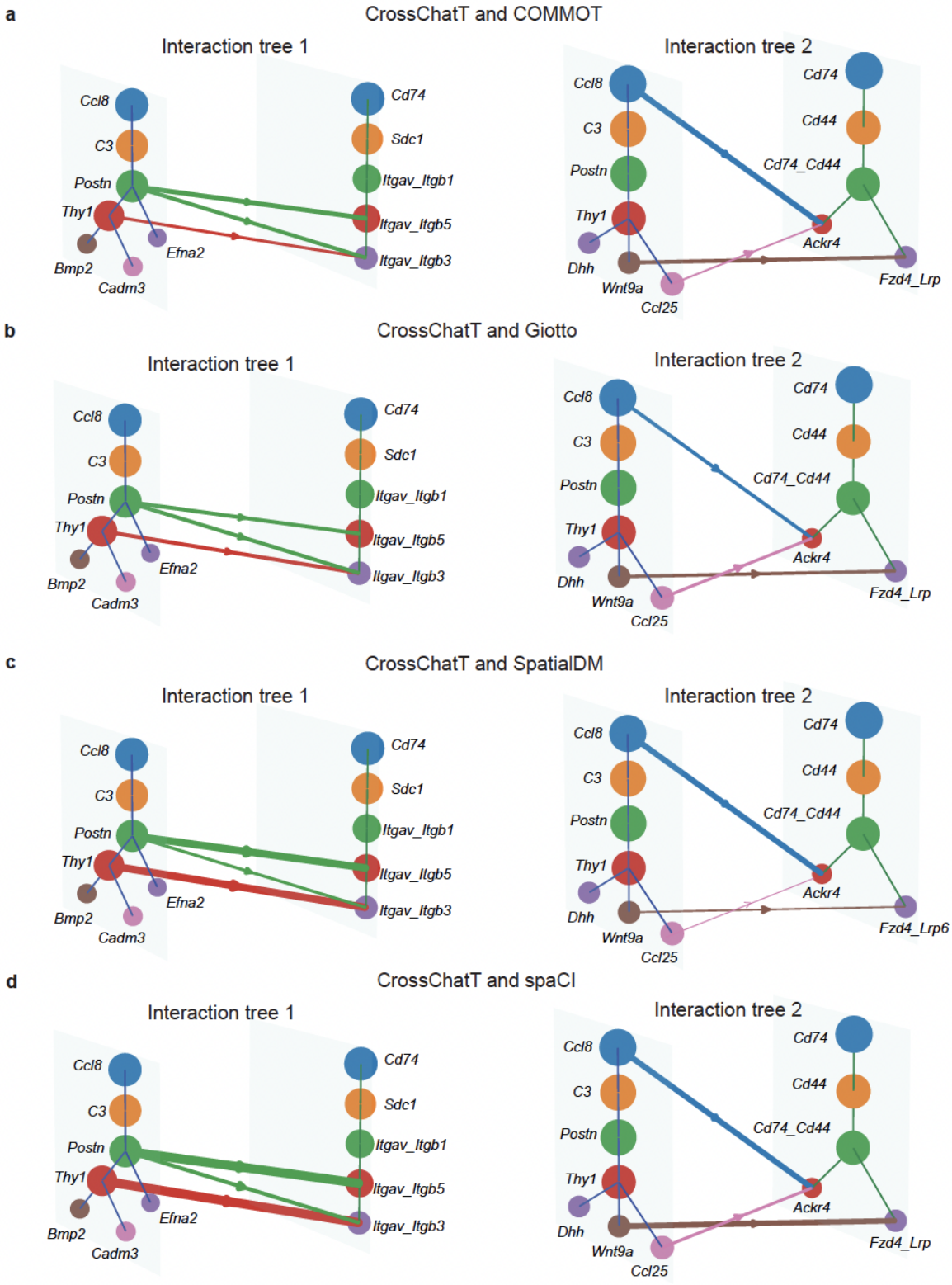


Figure C.9: **Comparison of CrossChatT analysis with different CCC methods.**

Figure C.9 (continued). (a) CCC detected by CrossChatT using COMMOT (b) CCC detected by CrossChatT using Giotto (c) CCC detected by CrossChatT using SpatialDM (d) CCC detected by CrossChatT using spaCI.

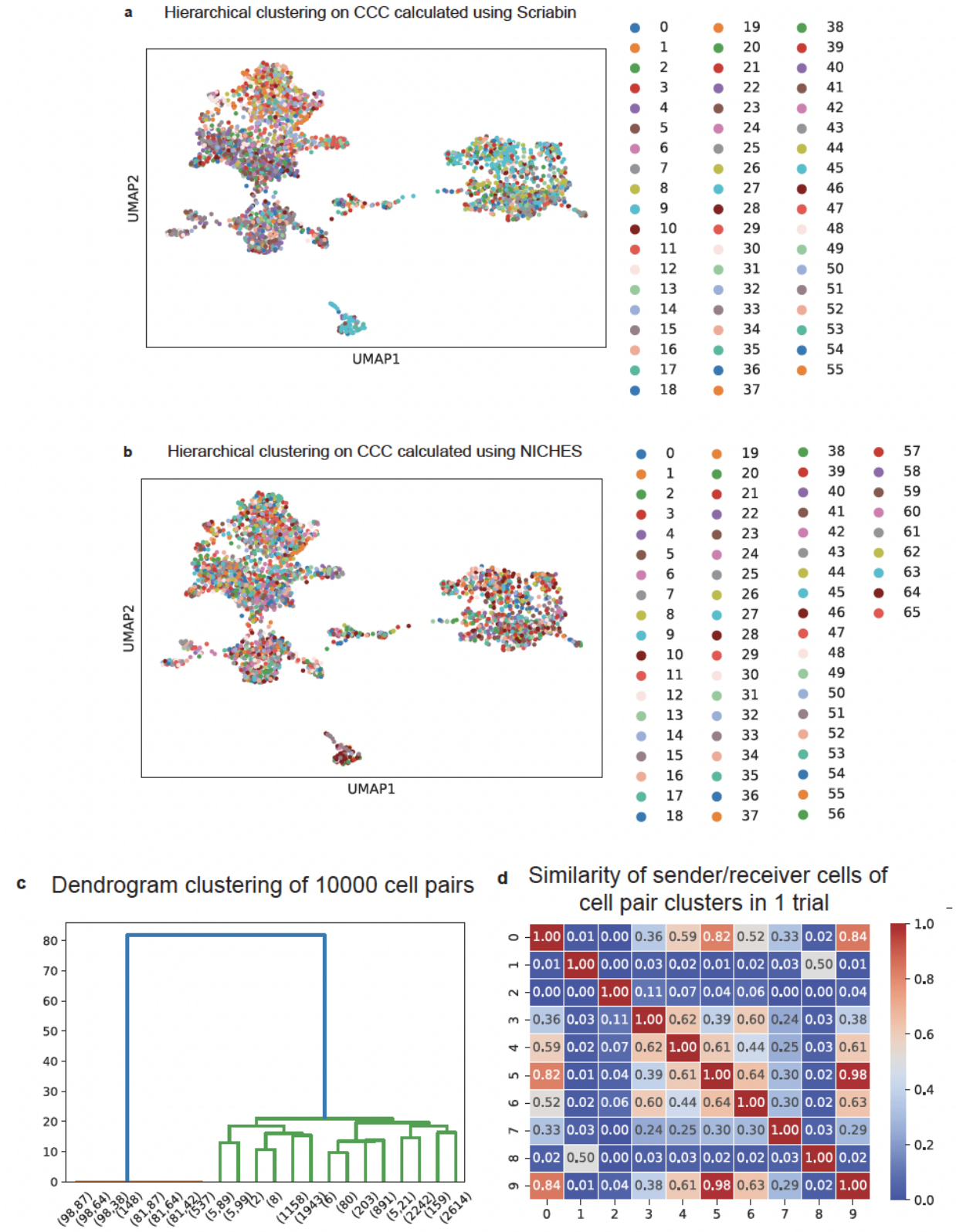


Figure C.10: **Hierarchical clustering on CCC calculated using single-cell level inference tools.**

Figure C.10 (continued). (a) CrossChatH clustering on CCC calculated using Scriabin (b) CrossChatH clustering on CCC calculated using NICHES (c) Dendrogram clustering of 10000 cell pairs of COVID-19 PBMC cells (d) Jaccard similarity of sender/receiver cells of cell pair clusters.