Comparative network analysis in the context of single-cell transcriptome-derived cell-cell communication

Von der Fakultät für Mathematik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

James Shiniti Nagai, M.Sc.

aus Mogi das Cruzes, São Paulo, Brazil

Berichter:

Universitätsprofessor Michael T. Schaub, PhD

Universitätsprofessor Dr. rer. nat. Ivan G. Costa

Tag der mündlichen Prüfung: *12.08.2026*

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

Dedicado para os meus Ditians e Batians
que sempre me encorajaram mesmo
em momentos tão difíceis.

Abstract (English)

Graphs are mathematical structures used to model relationships between entities and are widely applied to complex real-world data, including social, geographic, and molecular systems. In computational biology, advances in single-cell RNA sequencing (scRNA-seq) have led to extensive molecular datasets, where graphs are commonly used to represent and analyze biological processes in problems such as clustering and cell–cell communication (CCC).

Cell–cell communication, particularly through ligand–receptor interactions, plays a crucial role in tissue function and can be naturally modeled as directed, weighted graphs in which cells are nodes and interactions are edges. Although many tools exist to infer CCC from scRNA-seq data, comparing CCC graphs remains a challenging and an underexplored problem due to the complexity of directed and weighted graph structures, noise in real-world data, and limited interpretability of existing methods.

This thesis introduces new computational frameworks for comparing and analyzing CCC graphs. The contributions include: (1) CrossTalkeR, a method for prioritizing and comparing CCC graphs using node importance measures; (2) the development of optimal transport–based distance measures, including Wasserstein and Gromov–Wasserstein distances, for comparing directed weighted graphs; and (3) scACCorDiON, an optimal transport–based clustering approach for analyzing patient-specific CCC graphs in large cohorts. Methods are validated on real-world biological case studies on bone marrow fibrosis, myocardial infarction, COVID-19 and pancreatic cancer. These demonstrate that proposed approaches can improve interpretability, disease stratification, and biological relevance of cell-cell communication-driven mechanisms.

Abstrakt (Deutsch)

Graphen sind mathematische Strukturen, die zur Modellierung von Beziehungen zwischen Entitäten verwendet werden und in komplexen Daten aus der realen Welt, einschließlich sozialer, geografischer und molekularer Systeme, Anwendung finden. In der computergestützten Biologie haben Fortschritte in der Einzelzell-RNA-Sequenzierung (scRNA-seq) zu umfangreichen molekularen Datensätzen geführt, in denen Graphen häufig zur Darstellung und Analyse biologischer Prozesse in Problemen wie Clusteranalyse und Zell-Zell-Kommunikation verwendet werden.

Die Zell-Zell-Kommunikation, insbesondere durch Liganden-Rezeptor-Interaktionen, spielt eine entscheidende Rolle für die Gewebefunktion und kann auf natürliche Weise als gerichtete und gewichtete Graphen modelliert werden. Hierfür werden die Zellen durch Knoten und Interaktionen durch Kanten repräsentiert. Obwohl es viele Verfahren gibt, um Zell-Zell-Kommunikation aus scRNA-seq-Daten abzuleiten, bleibt der Vergleich von Zell-Zell-Kommunikationsgraphen aufgrund der Komplexität gerichteter und gewichteter Graphstrukturen, des Rauschens in realen Daten und der begrenzten Interpretierbarkeit bestehender Verfahren eine Herausforderung und ein noch wenig erforschtes Feld.

In dieser Doktorarbeit werden neue computerbasierte Frameworks zum Vergleich und zur Analyse von Zell-Zell-Kommunikationsgraphen vorgestellt. Diese umfassen: (1) CrossTalkeR, eine Methode zur Priorisierung und zum Vergleich von Zell-Zell-Kommunikations-Graphen unter Verwendung von Metriken zur Gewichtung der Knoten; (2) die Entwicklung von auf optimalem Transport basierenden Distanzmetriken, darunter Wasserstein- und Gromov-Wasserstein-Distanzen, zum Vergleich gerichteter gewichteter Graphen; und (3) scACCorDiON, eine auf optimalem Transport basierende Clusteranalyse für patientenspezifische Zell-Zell-Kommunikations-Graphen in großen Kohorten. Die Methoden werden anhand realer biologischer Fallstudien zu Knochenmarkfibrose, Myokardinfarkt, COVID-19 und Bauchspeicheldrüsenkrebs validiert. Diese zeigen, dass die vorgeschlagenen Ansätze die Interpretierbarkeit, die Krankheitsstratifizierung und die biologische Relevanz von Mechanismen, die durch die Zell-Zell-Kommunikation gesteuert werden, verbessern können.

The completion of this thesis would not have been possible without the support and assistance of my collaborators, friends, and family. I would like to begin this acknowledgment section by expressing my heartfelt gratitude to all of them.

Mãe, Pai (in memoriam) e Miko, vocês sempre me inspiram pra ser a minha melhor versão. Não existem palavras com que eu consiga expressar toda a minha gratidão e a felicidade de ter uma super família como a nossa.

Tio Americo, Tia Elza, Sayumi e Issao, a energia de ter vocês por perto é única! Sempre aprendo muito com vocês, desde as risadas nos finais de semana até o conforto nos dias mais difíceis. Muito obrigado por tornar todos os meus retornos mais especiais e confortáveis.

Prof. Dr. Ivan G. Costa and Prof. Dr. Michael T. Schaub, as I have mentioned in previous years, your support and warm mentoring have shaped my perspective and passion for research. I'm deeply grateful for having a safe space to express my science and for all the guidance, patience, and rich discussions.

Univ. Prof. Dr. med. Rebekka Schneider-Kramann and Prof. Dr. Rafael Kramann, I'm highly grateful for all the collaboration, awesome data, and discussions; tackling complex diseases with you is always fun and exciting!

Prof. Dr. Sikander Hayat, my gratitude for the cooperation and discussion we had on the projects.

A special thank you to all the first co-authors, particularly Dr. Jitske Jansen, Dr. Katharina Reimer, Dr. Helene Gleitz, Dr. Mayra Ruiz, Dr. Giulia Cesaro, and Bella Banjanin.

Ao meu amigos Julia Ribeiro, Lucas Vecchete e Felipe Asakawa, foram incontáveis as nossas aventuras, né? Todo o apoio nesta jornada fora do Brasil foi essencial. Um brinde à nossa amizade.

Ao meu irmão *tuguinha* Tiago Maie, que está ao meu lado desde aquele primeiro dia. Desde a primeira cerveja neste velho continente até às corridas e às inúmeras e enormes refeições que partilhamos. Com certeza, este árduo caminho, o doutoramento, foi mais divertido e transformador. Serei sempre grato e inspirado pela pessoa que você é.

Martin Manolov and Fabio Ticconi lost count of how many times I missed your names. Thank you for all the assistance, comments, and fun moments.

For my office mates, Dr. Mingbo Cheng and Dr. Martin Grasshof, there were so many occasions when we caught ourselves talking crazy computational biology methods on the whiteboard. I couldn't be happier with all that we have shared and learn with each other.

To my Iranian friends, Mina Shaigan and Mehdi Joodaki, I will always be thankful for all the culture and food you have both shared.

Bella Banjanin, what a journey!!! Have no words to say how happy I feel to have met you. Your passion for science and the way that you take care of yourself and others are things that will remain forever in my mind.

Zhijian Li and Ronghui Li, my eternal gratitude for all the discussions and guidance you all gave me. My journey with single-cell technologies wouldn't be that enjoyable without our discussions.

Vanessa Kloecker, sharing eMED meetings with you was awesome. I'm extremely lucky that I had you and your many skills throughout my PhD. From discussing CCC stuff to also fixing hoodies.

To Lelde, from the plant to all the rich Latvian culture. Thank you so much, my Latvian friend.

Let's go for food, Damin Kuehn? - Forever thankful for all the talks, discussions, science, and stuff

we did together, man!

To my mentees, Ralf, Martha, Laura, Vladimir, and Larissa. Mentoring you all was unique. I will always carry a bit of each of you with me throughout my life. I wish all of you the best. Extremely happy to have been sharing science with you.

I would also like to acknowledge the multiple reviewers of this thesis and manuscripts. All the inputs and corrections were extremely useful.

Ng Yin Fong, Mozi! Can't put into words how much you make my days better. I'm happy we shared so many things together, and that you are by my side.

Selbstständigkeitserklärung

Ich versichere hiermit an Eides statt, daß ich die vorliegende Doktorarbeit selbstständig und ohne unzuläßige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich habe die Grundsätze zur Sicherung guter wissenschaftlicher Praxis der RWTH Aachen zur Kenntnis genommen und eingehalten.

Aachen, August 17, 2026 _____

James Shiniti Nagai

Publications

As required by §5(3) of the

> *Promotionsordnung für die Fakultät für Mathematik, Informatik und Naturwissenschaften der Rheinisch-Westfälischen Technischen Hochschule Aachen vom 27.09.2010 in der Fassung der zweiten Ordnung zur änderung der Promotionsordnung vom 30.06.2014 (veröffentlicht als Gesamtfassung)*,

a declaration of results published by the author and particular contributions to co-authored publications follows.

The main authorship publications and the main content of this thesis were related to the following publications:

- **Nagai, J. S.**, Leimkühler, N. B., Schaub, M. T., Schneider, R. K., & Costa, I. G., 2021, CrossTalkeR: analysis and visualization of ligand–receptor networks, Bioinformatics, 37(22), 4263–4265.
  doi: 10.1093/bioinformatics/btab370.

- Jansen, J.; Reimer, K. C.; **Nagai, J. S.**; Varghese, F. S.; Overheul, G. J.; de Beer, M.; Roverts, R.; Daviran, D.; Fermin, L. A. S.; Willemsen, B.; Beukenboom, M.; Djudjaj, S.; von Stillfried, S.; van Eijk, L. E.; Mastik, M.; Bulthuis, M.; Dunnen, W. D.; van Goor, H.; Hillebrands, J.-L.; ... Kramann, R., 2022, SARS-CoV-2 infects the human kidney and drives fibrosis in kidney organoids, Cell Stem Cell, 29(2), 217–231.e8.
  doi: 10.1016/j.stem.2021.12.010.

- **Nagai, J. S.**, Costa, I. G., & Schaub, M. T., 2024, Optimal transport distances for directed, weighted graphs: a case study with cell-cell communication networks, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024), pp. 9856–9860.
  doi: 10.1109/ICASSP48485.2024.10446503.

- **Nagai, J. S.**, Maié, T., Schaub, M. T., & Costa, I. G., 2025, scACCorDiON: a clustering approach for explainable patient level cell–cell communication graph analysis, Bioinformatics, 41(5), btaf288.
  doi: 10.1093/bioinformatics/btaf288.

- Banjanin, B.; **Nagai, J. S.**; Mun, Y.; Fuchs, S.; Nombela-Arrieta, C.; Costa, I. G.; Schneider, R. K., 2025, Wnt-dependent spatiotemporal reprogramming of bone marrow niches drives fibrosis, Hemasphere, 10 no. 2 (2026): e70309. doi: 10.1002/hem3.70309.

I was also able to be part of two literature review papers:

- Dimitrov, D.; Türei, D.; Garrido-Rodriguez, M.; Burmedi, P. L.; **Nagai, J. S.**; Boys, C.; Ramirez Flores, R. O.; Kim, H.; Szalai, B.; Costa, I. G.; Valdeolivas, A., 2022, Comparison of methods and resources for cell–cell communication inference from single-cell RNA-Seq data, Nature Communications, 13(1), 3224.
  doi: 10.1038/s41467-022-30755-0.

- Cesaro, G.; **Nagai, J. S.;** Gnoato, N.; Chiodi, A.; Tussardi, G.; Klöker, V.; Musumarra, C. V.; Mosca, E.; Costa, I. G.; Di Camillo, B.; Calura, E.; Baruzzo, G., 2025, Advances and challenges in cell–cell communication inference: a comprehensive review of tools, resources, and future directions, Briefings in Bioinformatics, 26(3), bbaf280.
  doi: 10.1093/bib/bbaf280.

Additionally, I contributed to cell-cell communication analysis and/or served as the main bioinformatician in the following publications. Although those were not the main part of the thesis, they were highly valuable for understanding and learning about complex traits in single-cell biology.

- Leimkühler, N. B.; Gleitz, H. F. E.; Ronghui, L.; Snoeren, I. A. M.; Fuchs, S. N. R.; **Nagai, J. S.**; Banjanin, B.; Lam, K. H.; Vogl, T.; Kuppe, C.; Stalmann, U. S. A.; Büsche, G.; Kreipe, H.; Gütgemann, I.; Krebs, P.; Banz, Y.; Boor, P.; Tai, E. W. Y.; Brümmendorf, T. H.; … Schneider, R. K., 2021, Heterogeneous bone-marrow stromal progenitors drive myelofibrosis via a druggable alarmin axis, Cell Stem Cell, 28(4), 637–652.e8.
  doi: 10.1016/j.stem.2020.11.004.

- Peisker, F.; Halder, M.; **Nagai, J. S.**; Ziegler, S.; Kaesler, N.; Hoeft, K.; Li, R.; Bindels, E. M. J.; Kuppe, C.; Moellmann, J.; Lehrke, M.; Stoppe, C.; Schaub, M. T.; Schneider, R. K.; Costa, I. G.; Kramann, R., 2022, Mapping the cardiac vascular niche in heart failure, Nature Communications, 13(1), 3027.
  doi: 10.1038/s41467-022-30682-0.

- Jansen, J.; van den Berge, B. T.; van den Broek, M.; Maas, R. J.; Daviran, D.; Willemsen, B.; Roverts, R.; van der Kruit, M.; Kuppe, C.; Reimer, K. C.; Di Giovanni, G.; Mooren, F.; Nlandu, Q.; Mudde, H.; Wetzels, R.; den Braanker, D.; Parr, N.; **Nagai, J. S.**; Drenic, V.; … Smeets, B., 2022, Human pluripotent stem cell-derived kidney organoids for personalized congenital and idiopathic nephrotic syndrome modeling, Development, 149(9), dev200198.
  doi: 10.1242/dev.200198.

- Kaesler, N.; Cheng, M.; **Nagai, J. S.**; O'Sullivan, J.; Peisker, F.; Bindels, E. M. J.; Babler, A.; Moellmann, J.; Droste, P.; Franciosa, G.; Dugourd, A.; Saez-Rodriguez, J.; Neuss, S.; Lehrke, M.; Boor, P.; Goettsch, C.; Olsen, J. V.; Speer, T.; Lu, T.-S.; Lim, K.; Floege, J.; Denby,

L.; Costa, I. G.; Kramann, R., 2023, Mapping cardiac remodeling in chronic kidney disease, Science Advances, 9(47), eadj4846.
doi: 10.1126/sciadv.adj4846.

- Miguel, V.; Rey-Serra, C.; Tituaña, J.; Sirera, B.; Alcalde-Estévez, E.; Herrero, J. I.; Ranz, I.; Fernández, L.; Castillo, C.; Sevilla, L.; Nagai, J. S.; Reimer, K. C.; Jansen, J.; Kramann, R.; Costa, I. G.; Castro, A.; Sancho, D.; Rodríguez González-Moro, J. M.; Lamas, S., 2023, Enhanced fatty acid oxidation through metformin and baicalin as therapy for COVID-19 and associated inflammatory states in lung and kidney, Redox Biology, 68, 102957.
doi: 10.1016/j.redox.2023.102957.

- Pritchard, J. E.; Pearce, J. E.; Snoeren, I. A. M.; Fuchs, S. N. R.; Götz, K.; Peisker, F.; Wagner, S.; Benabid, A.; Lutterbach, N.; Klöker, V.; **Nagai, J. S.**; Hannani, M. T.; Galyga, A. K.; Sistemich, E.; Banjanin, B.; Flosdorf, N.; Bindels, E.; Olschok, K.; Biaesch, K.; Chatain, N.; Bhagwat, N.; Dunbar, A.; Sarkis, R.; Naveiras, O.; Berres, M.-L.; Koschmieder, S.; Levine, R. L.; Costa, I. G.; … Schneider, R. K., 2024, Non-canonical Hedgehog signaling mediates profibrotic hematopoiesis-stroma crosstalk in myeloproliferative neoplasms, Cell Reports, 43(1), 113608.
doi: 10.1016/j.celrep.2023.113608.

- Schlößer, S.; Ullrich, A.-L.; Fazel Modares, N.; Schmitz, M. A.; Schöneich, J.; Zhang, K.; Richter, I.; Robrahn, L.; Schraven, S.; **Nagai, J. S.**; Haange, S.-B.; Jennings, S. A. V.; Clavel, T.; Rolle-Kampczyk, U.; Kiessling, F.; Costa, I. G.; Muncan, V.; Repnik, U.; von Bergen, M.; Dupont, A.; Hornef, M. W., 2024, Salmonella infection accelerates postnatal maturation of the intestinal epithelium, Proceedings of the National Academy of Sciences of the United States of America, 122(1), e2403344122.
doi: 10.1073/pnas.2403344122.

- Gleitz, H. F. E.; Fuchs, S. N. R.; Snoeren, I. A. M.; Boys, C.; Nagai, J. S.; Tejeda-Mora, H.; Klöker, V.; Pritchard, J. E.; Bakker, I. J.; Gargallo Garasa, M.; Bindels, E.; Saez-Rodriguez, J.; Vogl, T.; Kramann, R.; Dugourd, A.; Costa, I. G.; Schneider, R. K., 2025, Inhibiting the alarmin-driven hematopoiesis-stromal cell crosstalk in primary myelofibrosis ameliorates bone marrow fibrosis, HemaSphere, 9(8), e70179.
doi: 10.1002/hem3.70179.

Finally, I contributed to several other computational biology tools as a co-author.

- Li, Zhijian, **James S. Nagai**, Christoph Kuppe, Rafael Kramann, and Ivan G. Costa. "scMEGA: single-cell multi-omic enhancer-based gene regulatory network inference." **Bioinformatics Advances** 3, no. 1 (2023): vbad003.

- Joodaki, Mehdi, Mina Shaigan, Victor Parra, Roman D. Bülow, Christoph Kuppe, David L. Hölscher, Mingbo Cheng, **James Nagai** et al. "Detection of PatIent-Level distances from single cell genomics and pathomics data with Optimal Transport (PILOT)." **Molecular Systems Biology** 20, no. 2 (2024): 57-74.

- Mehdi Joodaki, Mina Shaigan, Samaneh Samiei, **James Nagai**, Tiago Maié, Christoph Kuppe, Ivan G Costa, PILOT-GM-VAE: patient-level analysis of single-cell disease atlas with optimal transport of Gaussian mixture variational autoencoders, Briefings in Bioinformatics, Volume 26, Issue 5, September 2025, bbaf547, https://doi.org/10.1093/bib/bbaf547

- Cheng, M., Jansen, J., Reimer, K.C., Grande, V.P., Nagai, J.S., Li, Z., Kießling, P., Grasshoff, M., Kuppe, C., Schaub, M.T. and Kramann, R., 2025. PHLOWER leverages single-cell multimodal data to infer complex, multi-branching cell differentiation trajectories. Nature Methods, pp.1-9.

We refer to the records in Google Scholar for other collaboration papers performed during the doctoral studies and generated with the support of eMED Fibromap Consortia.

To honor the fact that research is rarely an entirely solitary process, I will, in this thesis, rely on the use of the first-person plural pronoun “we” in the text as a nosism.

List of Figures

1.1 From Real world data to Graph Representation 2
1.2 Summary of the thesis main contributions of this thesis. 4

2.1 Real-world applications for graph theory. 6
2.2 Common node importance measures present in the literature 7
2.3 Learning distances across graphs 8
2.4 Graph Edit Distance Problem 10
2.5 Scheme of Optimal Transport 10
2.6 Gromov-Wasserstein Scheme 12
2.7 Cell-Cell Communication within a Tissue 13
2.8 Ligand-Receptor cell-cell communication 13
2.9 Single-cell RNA sequencing protocol 14
2.10 Identification of cell types using single-cell RNA sequencing 15
2.11 Cell-Cell Communication inference using scRNAseq data 16

3.1 From LR interactions to Graphs: 22
3.2 CrossTalkeR scheme 25
3.3 From LR inference results to graphs 26
3.4 Fibrosis formation example in the context of wound healing. Adapted from Younesi et al. (2024) 32
3.5 scRNAseq from Bone Marrow biopsy samples. 33
3.6 Cell-Cell Communication from BM fibrosis assembled by CrossTalkeR. 34
3.7 CrossTalkeR downstream analysis output 34
3.8 Megakaryocyte top 100 interactions 35
3.9 Myeloid top 100 interactions 36
3.10 Revealing S100A8/S100A9 as fibrosis signatures 36
3.11 COVID and Kidney Fibrosis scRNAseq and TAPseq 37
3.12 Investigating the connection between COVID and Kidney Fibrosis 38

4.1 Graph Kernel × Graph Embeddings 42
4.2 Rationale behind the grounded Laplacian 43
4.3 Magnetic Laplacian direction encoding 44
4.4 Undirected graph and node-to-node distance measures 45
4.5 Scheme for Shared Topology Graph 48

4.6 Synthetic Directed Weighted Graph examples 49
4.7 Synthetic Data Results 1 50
4.8 Synthetic Data Results 2 51
4.9 Synthetic Data Results 3 51
4.10 Topological variables per perturbation 52

5.1 Overview of scACCorDiON: 61
5.2 Patient CCC Clustering Benchmark 65
5.3 Patient CCC 2D Embeddings 66
5.4 Sub-cluster analysis in pancreas adenocarcinoma 68
5.5 Optimal Transport mapped features per sample 69
5.6 Top 20 ligand-receptor pairs Ductal cells 70
5.7 Pipeline of LR validation 71

A.1 Flowchart with the experimental design of scACCorDiON. 93
A.2 Patient Clustering Benchmark 94
A.3 RI benchmark 95
A.4 Silhouette analysis for scRNAseq patient cohorts 95
A.5 Downstream analysis using Tensor-cell2cell 96
A.6 PDAC Functional Analysis 97

List of Tables

3.1 CrossTalkeR input format 26

4.1 K-Medoids Adjusted Rand Index 53
4.2 KNN Classifier Mean Accuracy 53

5.1 Benchmark Dataset Description 63
5.2 Adjusted Random Index per dataset 65
5.3 Results from cluster resolution assay 67
5.4 Survival Analysis hits using the top 10 Malignant Ductal->Ductal cell in PDAC 2 . . 71
5.5 Survival Analysis hits using the top 10 Malignant Ductal->Ductal cell in PDAC 1 . . 72
5.6 Survival Analysis hits using the top 10 Malignant Ductal->Ductal cell in Factor4 . . 73
5.7 Survival Analysis hits using the top 10 Malignant Ductal->Ductal cell in Factor1 . . 73

6.1 Example of CrossTalkeR input in a urban mobility context 77

Glossary

AG	Agonist
AN	Antagonist
ARI	Adjusted Rand Index
BM	Bone Marrow
CCC	Cell-Cell Communication
CCI	Cell-Cell Interaction
CCN	Cell-Cell Network
CGN	Cell-Gene Network
COVID	Coronavirus Disease
CPDB	CellPhoneDB
CTR	Control
DNA	Deoxyribonucleic acid
DWG	Directed Weighted Graph
ECM	Extracellular Matrix
EXP	Experiment
Fib	Fibroblast
GED	Graph Edit Distance
GOT	Graph Optimal Transport
GRD	Generalized Resistance Distance
GWOT	Gromov-Wasserstein Optimal Transport
HTD	Hitting Time Distance
kNN	k-Nearest Neighbors
LLM	Large Language Models
LR	Ligand-Receptor
MK	Megakaryocyte
MPNs	Myeloproliferative neoplasms

MSC	Mesenchymal Stem Cells
NATMI	Network Analysis Toolkit for Multicellular Interactions
OT	Optimal Transport
PCA	Principal Component Analysis
PDAC	Pancreatic adenocarcinoma
PMF	Primary Myelofibrosis
PPI	Protein-Protein Interactions
PT	Proximal Tubule
RI	Rand Index
RNA	Ribonucleic acid
SBM	Stochastic Block Models
scRNAseq	single-cell RNA sequencing
ST	Spatial Transcriptomics

Contents

Abstract (English) . . . i
Abstrakt (Deutsch) . . . iii
Acknowledgements . . . v
Selbstständigkeitserklärung (Declaration of Authenticity) . . . vii
Publications . . . xii
List of Figures . . . xiv
List of Tables . . . xv
Glossary . . . xvii

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Contributions . . . 2
1.3 Thesis Structure . . . 3

2 Background **5**
2.1 Notation . . . 5
2.2 Graph Theory . . . 5
2.2.1 Node importance . . . 7
2.2.2 Learning Graph Distances . . . 9
2.3 Introduction to Cell-Cell Communication . . . 12
2.3.1 Cell-Cell Communication in terms of biology . . . 12
2.3.2 Single Cell RNA sequencing (scRNAseq) . . . 13
2.4 Inference of cell-cell communication from scRNAseq data . . . 16
2.5 Discussion . . . 19

3 Comparing pairs of cell-cell communication graphs **21**
3.1 Motivation and Related Work . . . 21
3.1.1 CellChat . . . 22
3.1.2 NATMI . . . 23
3.1.3 Limitations of previous work . . . 24
3.2 CrossTalkeR . . . 24
3.2.1 Network Construction . . . 25
3.2.2 Differential Network Analysis . . . 27
3.2.3 Statistical test for comparing CCC . . . 29
3.2.4 CrossTalkeR technical requirements . . . 31

3.2.5 Current state of CrossTalkeR and Next Features . . . 31
3.3 Applications using CrossTalkeR in Fibrotic Diseases . . . 31
3.3.1 Fibrosis . . . 31
3.4 Discussion . . . 39

4 Learning distances from a set of cell-cell communication graphs 41
4.1 Motivation and Related Work . . . 41
4.1.1 Undirected graph optimal transport . . . 42
4.1.2 Graph Laplacian in Directed Weighted Graphs . . . 43
4.1.3 Graph node-to-node distances . . . 44
4.2 Learning metrics using directed weighted graphs . . . 47
4.2.1 Directed Weighted Graph Wasserstein Distance . . . 47
4.2.2 Directed Weighted Graph Gromov-Wasserstein Distance . . . 48
4.2.3 Synthetic model for benchmarking distances in graphs . . . 49
4.2.4 Benchmark strategies . . . 50
4.3 Discussion . . . 54

5 Clustering of cell-cell communication graphs 57
5.1 Motivation and Related Work . . . 57
5.1.1 Tensor-Cell2Cell . . . 57
5.1.2 MultiNicheNet . . . 59
5.1.3 Limitations of Current Tools . . . 59
5.2 Clustering Methods via Expectation Maximization . . . 60
5.2.1 Wasserstein Barycenters . . . 60
5.3 scACCorDiON: Clustering Patient Cell-Cell Communication Graphs . . . 61
5.3.1 Clustering Patients' Networks . . . 62
5.4 Benchmark . . . 63
5.4.1 Benchmarking cell-cell communication graph clustering . . . 64
5.4.2 Robustness analysis . . . 67
5.4.3 Case study: Pancreatic Ductal Adenocarcinoma . . . 68
5.5 Discussion . . . 72

6 Discussions and Future Work 75
6.1 Overall Discussion . . . 75
6.1.1 Differential Graph Analysis . . . 75
6.1.2 Sample Level Graph Analysis . . . 76
6.1.3 Beyond Cell-Cell Communication . . . 77
6.2 Future Work . . . 77
6.2.1 Spatial Transcriptomics and Cell-Cell Communication . . . 78
6.2.2 Beyond Graphs and Multimodal Cell-Cell Communication . . . 78

A Appendix **91**
A.1 MultiNicheNet 91
A.2 Unbalanced Optimal Transport 92

CHAPTER 1

Introduction

1.1 Motivation

Graphs are mathematical objects that are used to represent relationships between variables. In their primitive form, graphs encode the interactions (edges) between actors (nodes). Currently, graphs are widely employed to model/analyze an unlimited number of complex real-world systems such as social networks (Freeman, 1977), geographical data (Furutani et al., 2019), and molecular data (Hou et al., 2020; Jin et al., 2021). In this diverse context, graph comparison is a task that helps identify critical changes in variables in complex systems. Besides the meaningful value of methods comparing graphs, it is a hard task, given its combinatorial nature.

In computational biology and bioinformatics, advances in high-throughput sequencing technologies have led to an extensive reservoir of molecular data (CZI Single-Cell Biology et al., 2023) and have raised new research questions in recent years. Specifically, single-cell RNA sequencing (scRNA-seq) has been used in numerous studies to investigate complex biological processes, including diseases and human development (CZI Single-Cell Biology et al., 2023). Within current best practices (Lähnemann et al., 2020), graphs are widely used across several steps, including single-cell clustering, rare-cell detection, and cell-cell communication (CCC) analysis.

Cell-cell communication (CCC) occurs in various ways and dictates how tissues within an organism function. Ligand-receptor (LR) interactions are among the main interaction types that govern CCC (Armingol et al., 2024; Cesaro et al., 2025). As illustrated in Figure 1.1, a source cell communicates with a target cell by releasing signaling molecules to the environment, which are recognized by a cognate receptor in the target cell. Such interactions can be easily represented as directed graphs, where cells are actors, directed edges are oriented from ligand to receptor, and edge weights represent the number of molecules exchanged in this specific cell-cell interaction.

With more than a hundred tools published in recent years, scRNA-seq-derived CCC is a rapidly growing field of research (Armingol et al., 2024; Cesaro et al., 2025; Cheng et al., 2025). To infer CCC, these methods require the scRNA-seq count matrix and prior knowledge of communication. The prior knowledge of communication is a list of LR pairs: a ligand is a gene associated with biological compounds that a cell releases into the environment, and a receptor is a gene usually associated with a cell membrane protein that binds those ligands. Note that some cell-cell interactions occur in an autocrine/intracrine manner, i.e., the cell interacts with itself, and the receptors are not necessarily located at the cell membrane. A list of the relevant communication interactions is generated as the output of these inferences. Although communication interactions can naturally be represented as

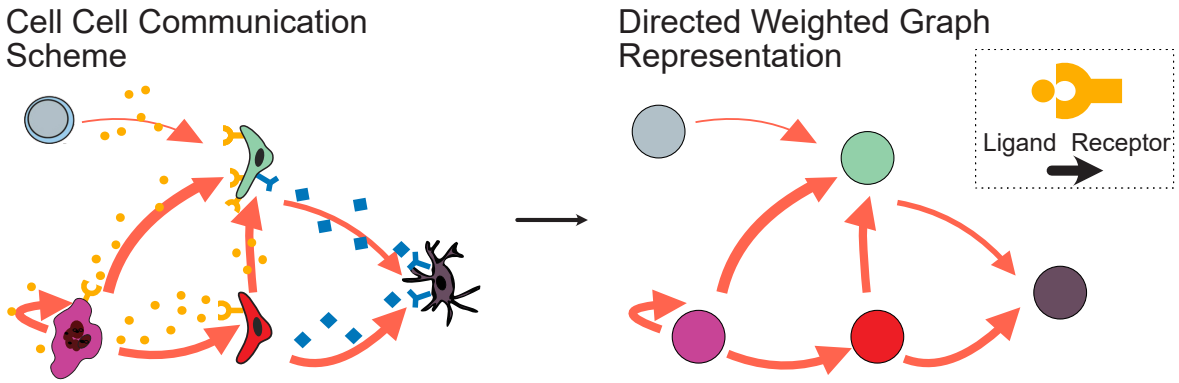


Figure 1.1: Cell-Cell Communication in nature yields a directed graph representation. LR pairs (ligands are represented by yellow circles and blue squares) indicate the cell-cell interaction direction. The thickness of each interaction is proportional to the number of molecules associated with the cells within the interaction.

graphs, CCC graphs remain an open and exciting research field.

In computational terms, signed directed graphs, such as CCC graphs, pose several challenges due to their non-linear and asymmetric nature (Diaz-Diaz et al., 2025). Furthermore, many developed methods neither account for the true nature of real-world data (e.g., noise levels and interpretability) nor are accessible to non-computational researchers.

1.2 Contributions

In this thesis, we (1) present a new computational model to prioritize cell-cell communication graphs and perform pairwise cell-cell communication graph comparisons, (2) we further develop an optimal transport-based method for learning and clustering distances using patient-derived cell-cell communication graphs, (3) we present real-world applications translating computational inferences to validation in real-world biological settings.

More specifically, our contributions are listed as follows:

- **CrossTalkeR, ranking and comparing pairwise cell-cell communication graphs:** While scRNAseq-derived CCC analysis often produces a ranked list of interactions, interpreting them is challenging. To facilitate such interpretation, we developed CrossTalkeR (Nagai et al., 2021), a computational framework that leverages CCC-directed graphs to compare and prioritize CCC events. CrossTalkeR implements a set of node importance measures that enable ranking of genes/cells important to CCC across single- and comparative-phenotype contexts. Although the tool initially focuses on the computational biology audience, this could also be translated to other interdisciplinary domains. CrossTalkeR results reveal previously known CCC mechanisms in Bone Marrow Fibrosis. Further use cases of CrossTalkeR encompass the identification of mechanisms in complex disease settings, such as in Bone Marrow Fibrosis, Myocardial infarction (Kuppe et al., 2022), and COVID Jansen et al. (2022).

- **Learning distances from directed weighted graphs using optimal transport:** Comparing

graphs using optimal transport has recently gained significant attention by providing a principled metric between graphs and an interpretable description of the associated changes between graphs in terms of a transport plan. In contrast to the computation of optimal transport for undirected graphs, the lack of symmetry introduces substantial challenges for implementing optimal transport for directed and weighted graphs. To this end, here, we repurposed and evaluated previously described node-to-node distances for directed and weighted graphs as the costs for the optimal transport. Further, we developed two distance measures to compare directed graphs based on variants of optimal transport: (i) an earth mover's distance (Wasserstein) and (ii) a Gromov-Wasserstein (GW) distance. We evaluate these two distances and discuss their relative performance for both simulated graph data and real-world directed cell-cell communication graphs inferred from single-cell RNA-seq data.

- **Clustering directed weighted graphs using optimal transport:** Current computational methods cannot yet analyze sample-specific cell-cell communication events, as measured in single-cell data produced in large patient cohorts. Cohort-based cell-cell communication analysis presents many challenges, such as the non-linear nature of cell-cell communication and the high variability given by the patient-specific single-cell RNA-seq datasets. Here, we present scACCorDiON (single-cell Analysis of Cell-Cell Communication in Disease clusters using Optimal transport in Directed Networks), an optimal transport-based clustering algorithm exploring node distances on the Markov Chain as the ground metric between directed weighted graphs. Benchmarking indicates that scACCorDiON performs better at clustering samples by disease status than competing methods that use undirected graphs. We present a case study of pancreatic adenocarcinoma in which scACCorDiON identifies a sub-cluster of disease samples characterized by changes in the tumor microenvironment. Our case study corroborates that clusters provide a robust and explainable representation of cell-cell communication events and that the expression of detected LR pairs is predictive of pancreatic cancer survival.

1.3 Thesis Structure

As illustrated in Figure 1.2, the thesis contains a chapter covering the key concepts underlying the main contributions of this work (Chapter 2). In Part I of this thesis, we will introduce the CCC comparison problem in the primary context (Chapter 3). Specifically, we use network topology to enable condition-based comparisons of how cells communicate in complex diseases. In Part II, we aim to extend the network comparison to cases where comparison of a set of networks is desired (Chapter 4 and Chapter 5). This cohort-wise view of the network comparison problem enables the identification of shared communication programs across patient subgroups.

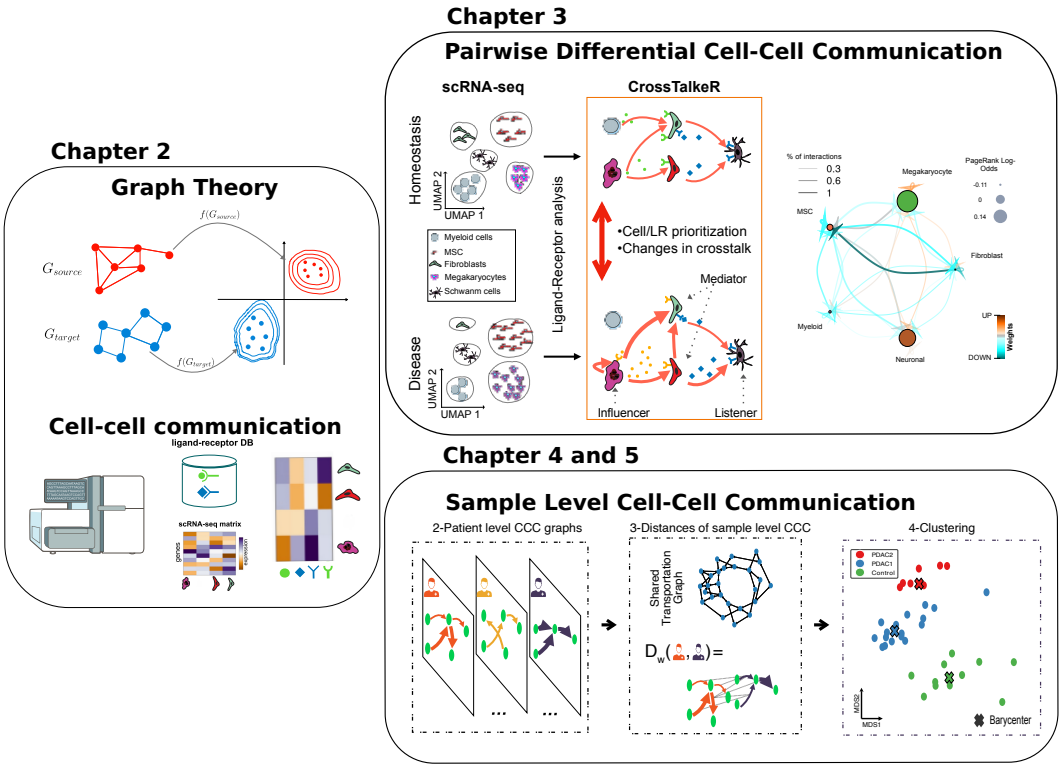


Figure 1.2: Summary of the main contributions of this thesis. This thesis is divided into two parts: Part I: A pairwise comparison view of CCC (Chapter 3), and Part II: Comparing many cell-cell communication networks (Chapter 4 and Chapter 5).

CHAPTER 2

Background

This chapter presents the background knowledge necessary to understand the thesis's content, encompassing computational and biological concepts. The chapter begins with the notation adopted throughout the thesis (Section 2.1). Next, we introduce graph theory concepts and open challenges in Section 2.2. To complement, we present Molecular biology, sequencing, and analysis of cell-cell communication (CCC) in Section 2.3. Section 2.4 shows how we can formalize CCC inference methods mathematically and presents well-established methods in the current literature. At the end, a discussion on the background concludes this chapter (Section 2.5).

2.1 Notation

- Matrices will be represented by boldface uppercase letters: $\mathbf{A}$
- Matrix of ones will be represented by boldface uppercase letters: $\mathbf{1}$
- Vectors will be represented by boldface lowercase letters: $\mathbf{a}$
- Scalars will be represented by cursive letters: a
- The i-th element of the vector $\mathbf{a}$ will be denoted by: $\mathbf{a}_i$
- The i-th row of the matrix $\mathbf{A}$ will be denoted by: $\mathbf{A}_{(i,:)}$
- The j-th column of the matrix $\mathbf{A}$ will be denoted by: $\mathbf{A}_{(:,j)}$
- The (i, j) element of the matrix $\mathbf{A}$ is defined as: $\mathbf{a}_{ij}$

2.2 Graph Theory

Graphs are mathematical objects that enable the modeling and study of relationships between or across observations. Real-world applications modeled by graphs range from the routing problem introduced by Euler (1741) and displayed in Figure 2.1 to recent advances in massive data knowledge graphs extracted from current large language models (LLMs) (Peng et al., 2024). Graphs provide a

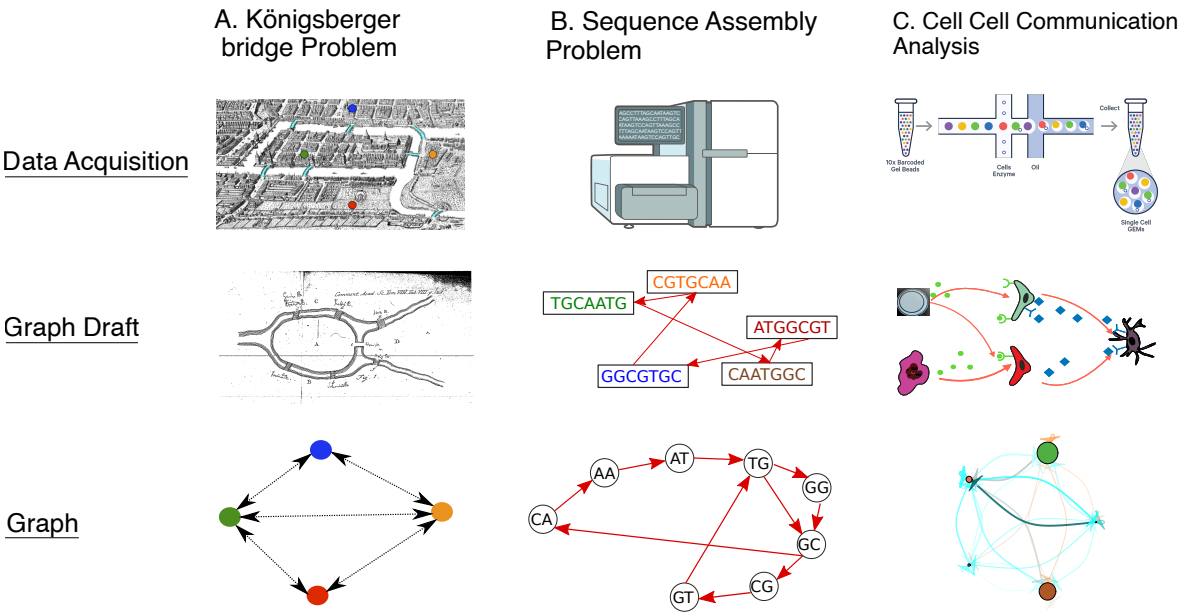


Figure 2.1: Real-world applications for graph theory. Shared areas refer to the different views used for data acquisition, synthetic modeling, and the final graph. *Adapted from:* www.10xgenomics.com, commons.wikimedia.org, bioart.niaid.nih.gov

general data representation which allows for use in several contexts, such as social networks (Freeman, 1977), urban traffic (Kim and Yoon, 2025), environmental science (Ishimoto et al., 2021), crop science (Aono et al., 2021), personalized medicine (Hou et al., 2020; Jin et al., 2021), and others. Even though there are several application areas for graphs, real-world problems often have complex features, which can further complicate simple graph models. Some examples of these complexities are: (1) the edge directionality as illustrated by the Königsberg bridge problem (Figure 2.1A); (2) the directed interactions in the genome assembly graphs (Figure 2.1B) and (3) the signed weights in the differential CCC graphs (Figure 2.1C).

While graphs can be represented in many complex ways, to better understand key concepts in graph theory, we now introduce and define the fundamental concepts of undirected graphs. Let an undirected graph $G = (V, E, W)$ where $e_i = (u_i, v_i) \in E$ is an edge connecting the nodes $u_i \in V$ and $v_i \in V$. $\mathbf{W}$ is the weighted adjacency matrix of the graph G

$$\mathbf{W} = \begin{cases} w_{uv} > 0 & \text{if } e_i = (u,v) \in \mathbb{E} \\ w_{uv} = 0 & \textit{otherwise.} \end{cases} \tag{2.1}$$

When we have a graph representation of the data, we can use the graph to extract node and edge statistics. An example of a node statistic is the node degree. In matrix notation, it is possible to calculate the node degree using the adjacency matrix $\mathbf{W}$. Another interesting tool that we can extract from graphs is the graph Laplacian. The graph Laplacian is a linear operator that captures how signals propagate through the graph. The graph Laplacian is of extreme importance in graph theory, being studied in both theoretical and applied settings. To compute the graph Laplacian, let $\mathbf{W}$ and $\mathbf{D} = diag(\mathbf{W1})$ be the weighted adjacency and degree matrices, respectively. We define the graph

Laplacian as $\mathbf{L} = \mathbf{D} - \mathbf{W}$ which can be represented in the matrix format as:

$$\mathbf{L} = diag\left(\overbrace{\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \ddots & \vdots & \\ 1 & 1 & \cdots & 1 \end{bmatrix}}^{\mathbf{D=W1}}\right) - \underbrace{\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix}}_{\mathbf{W}} \tag{2.2}$$

2.2.1 Node importance

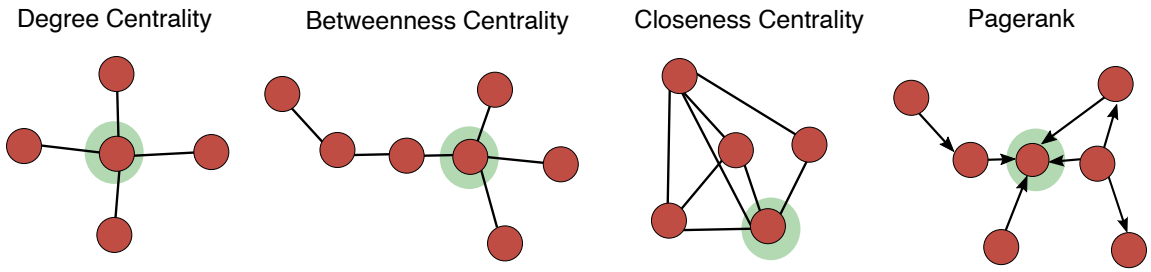


Figure 2.2: Common node importance measures present in the literature. Nodes represent observations, and the edges can be interpreted as a similarity measure (co-abundance, co-occurrence, or other).Highlighted nodes have the highest importance, as determined by the assigned value per measure. Each measure's rationale can be obtained using the following questions: How many observations am I connected to? (Degree Centrality); How many times am I bridging two other observations? (Betweenness Centrality); How fast am I able to reach all other observations? (Closeness Centrality); How important are the observations that I am interacting with? (PageRank centrality)

In graph theory, identifying important actors (nodes) given a graph is a primary task with numerous real-world applications. Introduced by Freeman (1977), centrality, or node importance, is a concept for identifying individuals with a central role within a graph.

Formally, a *node importance* (Centrality) measure of a node x within a graph G is defined as a mapping $\sigma^G : V_G \to \mathbb{R}_0^+$ which assigns a non-negative real value for each x $\in V_G$. Higher values of σ^G indicate greater interconnectedness (Landherr et al., 2010). *Node importance* measures are illustrated in Figure 2.2. The main examples of *node importance* measures are degree (Shaw, 1954), betweenness (Freeman, 1977), closeness, and the pagerank centrality (Page et al., 1998). Here, we provide definitions of measures widely used in several fields of the natural sciences.

Degree Centrality: The degree centrality in the undirected graphs is defined as $C_d(i) = D_{i,i}$. This measure relates the node's importance to the number of adjacent neighbors.

Betweenness Centrality: The betweenness centrality indicates that a node is important when it is an intermediate neighbor. Given u,v,x $\in V$, let τ_{uv} be the total number of the shortest paths between u and v and $\tau_{uv}(x)$ be the number of shortest paths that contain x. The Betweenness Centrality is computed as:

$$C_b(i) = \sum_{\substack{(u,v)\in E \\ u\neq v\neq x}} \frac{\tau_{uv}(x)}{\tau_{uv}}. \tag{2.3}$$

Closeness Centrality: The closeness centrality indicates how fast a node can communicate with other nodes in the graph. Let $d(u,v)$ denote a distance measure between the nodes u and v, the Closeness centrality is defined as:

$$C_{clo}(u) = \frac{1}{\sum_v^{|V|} d(u,v)}. \tag{2.4}$$

PageRank Centrality: Introduced by the pioneering work of Page et al. (1998), PageRank centrality is popularly used to rank nodes in several complex systems. To compute the PageRank of each node, a matrix containing each node's visitation probability must be calculated. This can easily be achieved by computing the random walk Laplacian $\mathbf{L_{rw}}$ as follows:

$$\mathbf{L_{rw}} = \mathbf{AD^{-1}}, \tag{2.5}$$

where $\mathbf{A}$ represents the adjacency matrix and $\mathbf{D^{-1}}$ represents the inverse of the degree matrix. Note that $\mathbf{D^{-1}}$ is defined when there are no isolated nodes in the graph.

Using a convex combination of the graph-based probability to visit a node and the overall probability of visiting a random node, one can compute the pagerank iteratively:

$$\pi^{t+1} = \alpha \mathbf{L_{rw}} \pi^t + (1-\alpha)\frac{1}{n}, \tag{2.6}$$

where n is the total number of nodes and α is the damping factor, which controls how the overall visiting probability is used to compute the PageRank. The iterative process stops when:

$$\|\pi^{(t+1)} - \pi^{(t)}\|_1 < \epsilon \text{ (for a small threshold such as } 1e^{-5}). \tag{2.7}$$

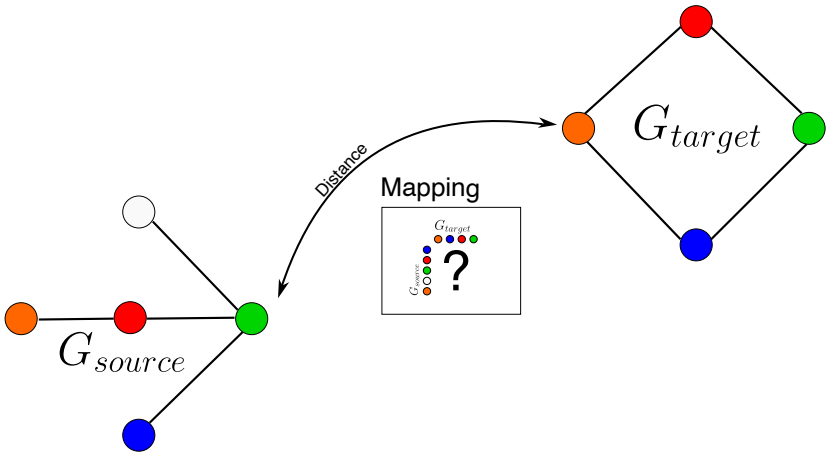


Figure 2.3: Learning distances across graphs: Node colors indicate a possible mapping between the graphs, which can aid in interpreting graph dissimilarities.

Studying node importance is a primary and complex research venue. With advances in modeling real-world problems using graphs, node importance measures can help clarify the role of an observation in an interaction context. Despite the well-established theory for undirected graphs presented above, representing real-world problems often demands complex models (e.g., signed graphs (Hou et al., 2020), directed weighted graphs (Hou et al., 2020; Jin et al., 2021), hypergraphs (Liu et al., 2025), high-order networks (Cheng et al., 2024), and others), which necessitates the development of new tailored measures for ranking nodes. In computational biology, centrality is also a widely used concept for identifying biomarkers of interest (e.g., disease-related genes, microorganisms, metabolic reactions, drugs, and others).

2.2.2 Learning Graph Distances

As mentioned above, node importance scores are a valuable proxy for ranking observations in a graph. While those measures can provide a score for each node, in most cases, they consider only a single graph. In a context where multiple graphs are observed/sampled from a given natural phenomenon (e.g., microbiology (Ishimoto et al., 2021)), understanding how graphs are similar can enable understanding shared properties across different graphs (Figure 2.3). Furthermore, identifying the major features of dissimilarity across graphs can provide an interpretable way to understand alterations in dynamic complex systems. The calculation of such dissimilarity can be achieved by a mapping across nodes in a graph as in Figure 2.3.

In graph theory, such a problem can be formulated as computing the graph edit distance (GED). In the graph edit distance problem, we seek to calculate the minimum number of changes (i.e., edit operations) required to transform a source graph into a target graph (Grohe and Schweitzer, 2020). In GED, only the following edit operations are allowed: (1) Change the label of edges or nodes; (2) Remove/Insert an edge between two existing nodes; (3) Remove or insert an isolated node (Xu and Chang, 2025). As the exact solution of GED is NP-hard, this problem is often solved using heuristics or time-consuming methods such as A* (Hart et al., 1968). In formal terms, GED is defined as follows:

Definition 2.1 (Graph Edit Distance)**.** *Let $G = (V_G, E_G)$ and $H = (V_H, E_H)$ be isomorphic undirected graphs. Then, there exists a one-to-one and an onto mapping (bijection) $f : V_G \to V_H$ s.t. for every pair of nodes $(u, v) \in V_G$, $(u, v) \in E_G$ if and only if $(f(u), f(v)) \in E_H$. Furthermore, there is a sequence path (edit path) of operations, node/edge insertion, node/edge deletion, and node/edge relabeling in which G is transformed to H. Within the previously mentioned context, the cost of an edit path is the sum of the costs associated with each edit operation in the path (see Figure 2.4). Finally, the graph edit distance $d_{GED}(G, H)$ is the minimum cost of an edit path that morphs G into H.*

Learning Graph Distances with Optimal Transport

Although the utility and extensive literature on Graph Edit Distance exists, computing such a distance poses computational challenges, primarily due to its combinatorial nature (Grohe and Schweitzer, 2020; Xu and Chang, 2025). Recently, optimal transport (OT) methods have been developed for

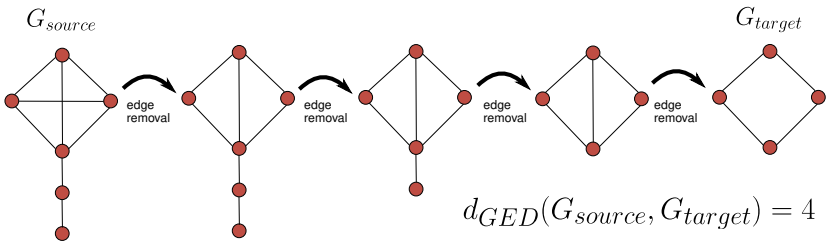


Figure 2.4: Example of a solution of the GED problem between G_{source} to G_{target}. Edit path between G_{source} and G_{target}. Edit operations are indicated at each step of the edit path.

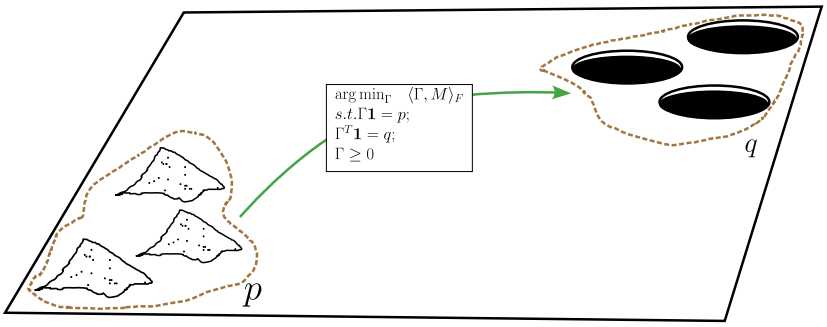


Figure 2.5: Illustration of optimal transport: The figure illustrates how optimal transport works: Given M, the distance between the holes and the piles of sand, OT computes the minimum effort to transport the pile of sand p to the holes q.

learning metrics on graphs. OT not only learns a metric between two graphs, but also generates a map that traces the changes. In terms of computational efficiency, computing such an optimization has time complexity $O(V^3)$ (e.g., GOT (Maretic et al., 2019)) which is comparable to the heuristics used to solve the GED optimization (Riesen and Bunke, 2009; Fischer et al., 2015). We next provide an introduction to OT, starting with its primary definition and followed by an example of OT methods for undirected graphs. OT is an approach developed by Gaspard Monge (Monge, 1781) to determine the minimal effort required to relocate a pile of soil to an excavation site, as illustrated in Figure 2.5. Formally, for discrete distributions, OT can be defined as the following minimization problem in Definition 2.2.

Definition 2.2 (Discrete Optimal Transport)**.** *Let p and q be probability distributions and a cost matrix M. The Wasserstein distance can be obtained by solving the following minimization problem.*

$$Wasserstein(p,q) = \arg\min_{\Gamma} \quad \langle \Gamma, M \rangle_F \tag{2.8}$$

$$s.t. \Gamma \mathbf{1} = p; \Gamma^T \mathbf{1} = q; \Gamma \geq 0. \tag{2.9}$$

In recent years, mainly motivated by the model explainability, OT (Figure 2.5) has been applied in several areas of natural sciences, such as computational biology (Schiebinger et al., 2019), image processing (Li et al., 2024), and others. In the context of graphs, a pioneering approach is called Graph Optimal Transport (GOT) (Maretic et al., 2019). GOT takes as input two undirected graphs and computes a distance measure based on their spectral properties. A detailed description is provided below.

Graph Optimal Transport (Maretic et al., 2019) extends the notion of OT for undirected graphs. In such an approach, graphs are considered key elements in which the probability distribution of signals is associated with the graph nodes. Precisely, let us consider two undirected graphs G_1 and G_2, with respective graph Laplacians L_1 and L_2. The signals of these graphs are, respectively, represented by:

$$\nu^{G_1} = \mathcal{N}(\mathbf{0}, L_1^{\dagger}) \tag{2.10}$$

$$\mu^{G_2} = \mathcal{N}(\mathbf{0}, L_2^{\dagger}), \tag{2.11}$$

where † denotes a pseudo-inverse operator.

Considering the signal distributions, a Wasserstein distance can be computed as follows:

$$W_2^2(\nu^{G_1}, \mu^{G_2}) = \inf_{T\#\nu^{G_1} = \mu^{G_2}} \int_{\chi} \|x - T(x)\|^2 d\nu^{G_1}(x), \tag{2.12}$$

where $T\#\nu^{G_1}$ represents the projection of ν^{G_1} by the transport map $T : \chi \to \chi$ defined on a metric space χ. When signals follow the normal distribution (Takatsu, 2010; Maretic et al., 2019), the Wasserstein distance and the respective transport map are defined by:

$$W_2^2(\nu^{G_1}, \mu^{G_2}) = Tr(L_1^{\dagger} + L_2^{\dagger}) - 2Tr(\sqrt{L_1^{\frac{\dagger}{2}} L_2^{\dagger} L_1^{\frac{\dagger}{2}}}), \tag{2.13}$$

and the respective OT transport map is $T(x) = L_1^{\frac{\dagger}{2}}(L_1^{\frac{\dagger}{2}} L_2^{\dagger} L_1^{\frac{\dagger}{2}}) L_1^{\frac{\dagger}{2}} x$.

Here, we only describe the Wasserstein distance for graphs introduced by GOT (Maretic et al., 2019). For the graph alignment optimization process, we refer to Maretic et al. (2019). An alternative perspective on the discrete optimal transport is the Gromov-Wasserstein OT (GWOT), as in Figure 2.6. In this optimization, instead of transporting probability distributions by considering a displacement between them, the transport accounts for displacements in the distances between them in two metric spaces. To be more precise, the GWOT is defined as follows:

Definition 2.3 (Gromov Wasserstein Optimal Transport)**.** *Let $\mathbb{X} = (X, d_x, \mu_x)$ and $(\mathbb{Y}, d_y, \mu_y)$ be two metric measure spaces in which d_x and d_y are distance metrics of $\mathbb{X}$ and $\mathbb{Y}$. $\mathbb{X}$ and $\mathbb{Y}$ possess also the probability distributions μ_x and μ_y. The Gromov-Wasserstein distance can then be obtained by solving the following minimization problem.*

$$GW(\mathbb{X}, \mathbb{Y}) = \min_{\Gamma \in \mathcal{T}} \sum L(d_x, d_y) T_x T_y, \tag{2.14}$$

where $T = \left\{\Gamma \mid \Gamma\mathbf{1} = \mu_x, \Gamma^T\mathbf{1} = \mu_y, \Gamma_{ij} \geq 0\right\}$ and L is a loss function that accounts for the differences between the distance metrics.

Recent applications of GWOT in graphs are possible by representing each graph as a metric space. Precisely each graph $G = (V, E, W)$ can be represented as $\mathbb{XG} = (X_G, d_{x_G}, \mu_{x_G})$ where X_G represents the nodes, d_{x_G} represents the node-to-node distances and μ_{x_G} a probability distribution in the node space. An example of such formulation in the context of graphs is described in Demetci et al. (2020)

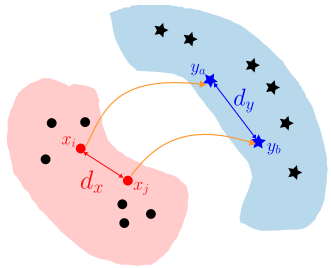


Figure 2.6: Illustration of the Gromov-Wasserstein Optimal Transport

2.3 Introduction to Cell-Cell Communication

In the current section, we introduce the context of this thesis's case study. To begin with, we situate the CCC in tissue biology. Next, we introduce the main mechanisms by which CCC can occur. Towards the end of the section, we introduce how scRNAseq can be performed and used to identify cell types.

2.3.1 Cell-Cell Communication in terms of biology

The human body is estimated to have approximately a trillion cells, which can be understood as units responsible for determining the function and structure of the body's organs (Bianconi et al., 2013). Inside the body, each organ function is characterized by a given set of cells that interacts with other cells and the tissue environment, as illustrated in Figure 2.7. Cell-cell interaction (CCI) is a key process in cell biology that orchestrates how organisms respond to various stimuli in their environment. Ligand-receptor interactions take place in specific CCI contexts, where source cells release proteins into the tissue microenvironment, called ligands. Ligands in the environment then bind to the cognate receptors in the target cells. As one can see, the interaction patterns are translatable into a graph structure where cells are the nodes and interactions are placed according to measurements of cognate molecules (e.g., combined gene expression). CCIs encompass a diverse range of molecules (e.g., junction proteins, ligands, receptors, and others) and can occur in the following four forms (Armingol et al., 2022).

- *Paracrine:* CCC occurs by propagating signalling molecules from one cell to another after secretion.
- *Endocrine:* Long-range CCC where secreted molecules travel long distances via extracellular fluids (e.g., blood plasma).
- *Autocrine:* Self-communication, ligands are secreted by cells, cognate receptors present within the same cell respond to the ligands and trigger a cellular response.
- *Juxtacrine:* Context-dependent CCC relies on specific tissue structures, such as gap junctions and membrane nanotubes. This communication type occurs without the secretion of molecules into the extracellular space.

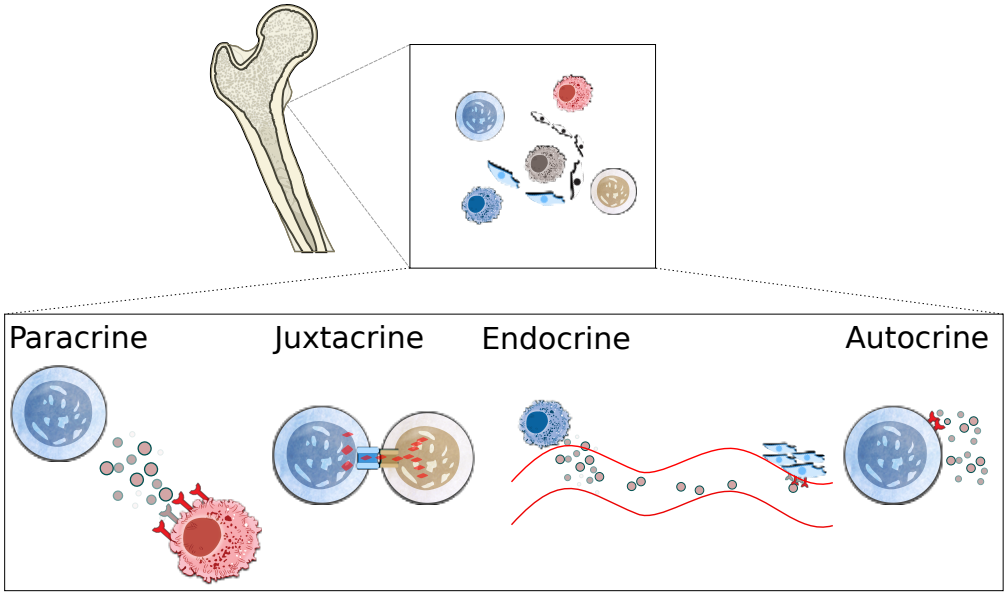


Figure 2.7: From tissues to CCC Example of bone marrow cells and how different ligand-receptor based cell-cell interaction types

Despite the diversity and importance of CCC for understanding complex diseases, measuring proteins directly requires tailored biochemical assays and extensive domain knowledge. Furthermore, studying specific proteins is challenging because they depend on their microenvironment/tissue. With the development of single-cell high-throughput sequencing technologies, the inference of CCC by using gene expression patterns as a proxy for communication molecules has successfully enabled a broader understanding of CCC patterns (Figure 2.8) in complex diseases (Armingol et al., 2022, 2024).

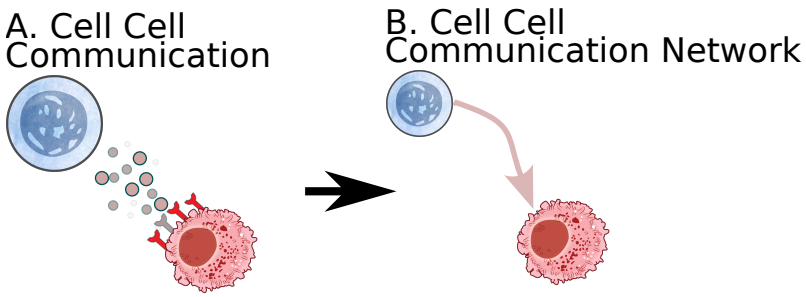


Figure 2.8: The nature of LR-mediated cell-cell communication yields a directed graph with a natural representation: nodes represent cells, and interactions are placed according to the LR orientation.

2.3.2 Single Cell RNA sequencing (scRNAseq)

RNA sequencing measures the gene expression profiles (transcriptomics) of a given biological sample (e.g., cancer biopsies, pest biopsies, or cell cultures) by acquiring, aligning, and counting RNA

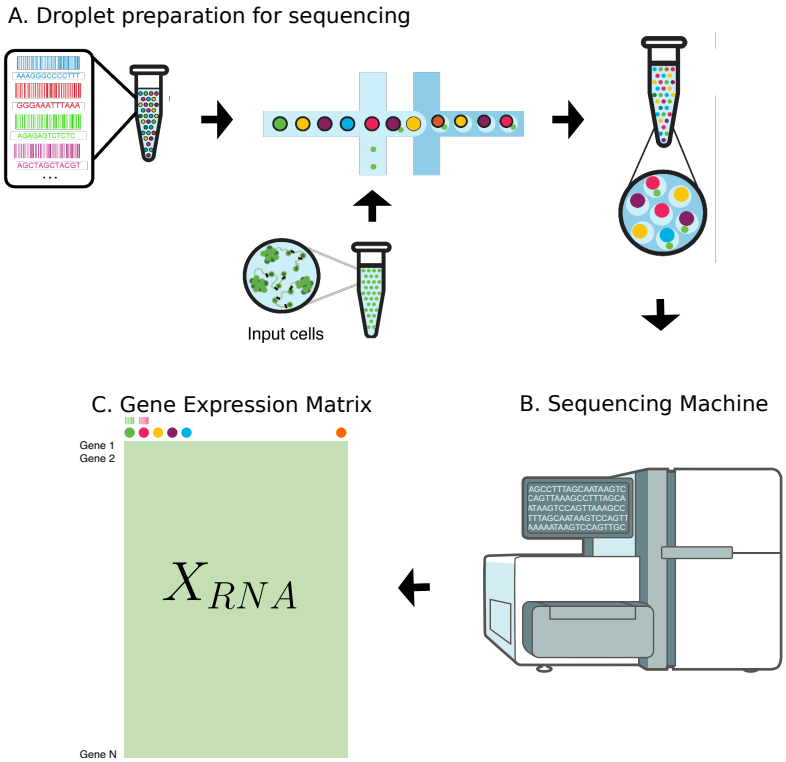


Figure 2.9: Droplet-based single-cell RNA sequencing: (From left to right) Barcode sequences are generated, cells are labeled with the barcodes, resulting in barcoded single-cell RNA sequencing droplets. Droplets are sequenced and aligned. This process generates a gene expression matrix that can be used to identify different cell populations.

sequences. Although the measurement allows for the quantification of RNA content in a sample, the sequencing process itself degrades the samples (Schiebinger et al., 2019). Such a limitation results in sampling challenges that must be accounted for when analyzing the data. With advances in sequencing technologies, significant amounts of single-cell-resolution sequencing data are being generated and used to expand knowledge of complex diseases. Acquiring single-cell transcriptomics enables a deeper understanding of key biological processes, such as aging and cancer. Although several protocols are currently available for single-cell sequencing, we will focus on microfluidic device-based protocols in this thesis. For a comprehensive and in-depth overview of current sequencing technologies, we refer to the Single-cell Best Practices resource https://www.sc-best-practices.org/.

Definition 2.4 (scRNAseq gene expression matrix)**.** *Let $X_{|g|,|sc|} \in \mathbb{R}$ be a matrix generated from a single-cell sequencing process; g and sc represent the set of genes and cells recovered in the sequencing, respectively. Each entry $X_{i,j} \geq 0$ represents the number of reads retrieved from gene i in cell j.*

As illustrated in Figure 2.9, in a manner analogous to how barcodes are used to identify products in retail stores, in microfluidic-based sequencing, barcodes (i.e., 16-base-pair DNA sequences) are used to identify each cell (Figure 2.9). The complementary DNA libraries from each cell's RNA are then fragmented, replicated, and sequenced. By aligning each DNA fragment to a reference genome, the number of RNA reads per gene is obtained. The sequencing process generates a matrix (Def. 2.4)

where genes are treated as features and cells as observations. Each entry of this matrix corresponds to the number of reads of a certain gene within a given cell. A common step in scRNAseq analysis is cell type identification (single-cell clustering). Such a step involves identifying clusters of cells with similar expression profiles. For performing the clustering, the following steps are needed: (a) Filtering; (b) Normalization; (c) Dimension Reduction.

(a) Filtering: Low-quality samples/features are removed from the data. Samples that present a large/small number of reads are disregarded, as well as the genes (features) with low expression frequency. Here, we are mainly interested in filtering cells with a low number of fragments or genes that are rarely expressed among the sequenced cells.

(b) Normalization: This step aims to remove the influence of technical effects in scRNAseq data, allowing cells and experiments to be more comparable. Some state-of-the-art normalization methods are scTransform (Hafemeister and Satija, 2019) and Counts per Million (Robinson et al., 2010) normalization.

(c) Dimensionality Reduction: A major challenge in scRNAseq data analysis is the high dimensionality of the data. While genes are quantified by the number of reads mapped to their genomic positions, the genes retrieved from sequencing vary across single cells. As a result, scRNAseq data exhibit a high dropout rate. A state-of-the-art approach to addressing the dropout issue is to use dimension-reduction methods (Lähnemann et al., 2020; Bouland et al., 2023). Standard scRNAseq pipelines use principal component analysis (PCA) of the gene expression matrix.

For clustering scRNAseq (Figure 2.10), a dimension-reduction matrix (e.g., PCA) is used to construct a cell-to-cell graph via k-nearest neighbors (kNN). A community-detection algorithm is applied to the kNN graph to generate single-cell clusters. Formally, we can define the scRNAseq clustering as presented in Definition 2.5

Definition 2.5 (scRNAseq clustering)**.** *Let* $sc = (sc_1, \ldots, sc_{|sc|})$ *be the set of single cells. The result of the scRNAseq clustering is a vector* $cl = (cl_1, \ldots, cl_{|sc|})$ *where* $cl_i \in \{1, 2, \ldots, k\}$ *with k being the total number of clusters and* $|sc|$ *is the total number of single cells.*

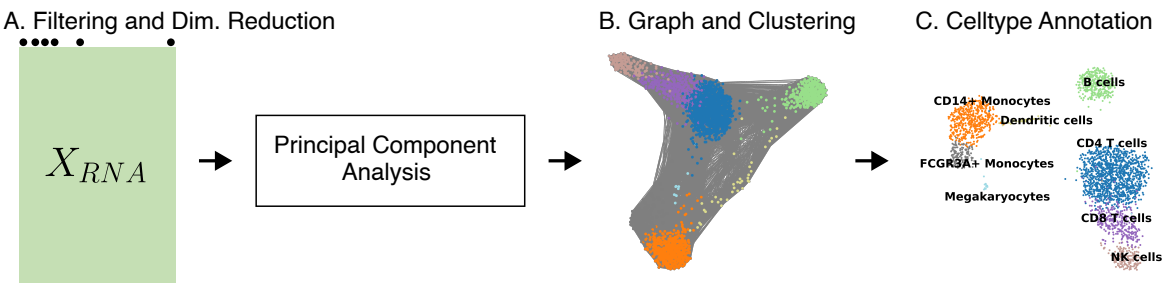


Figure 2.10: From the gene expression count matrix to the clustered single-cell graph

To provide a more precise explanation of the identified clusters, a statistical test is employed to identify specific gene signatures associated with each cluster. Finally, clusters are identified using

previously reported cell type signatures. Using the annotated scRNAseq dataset, several analyses can be performed to address a diverse range of biological questions from a gene-expression perspective; for example, CCC inference.

2.4 Inference of cell-cell communication from scRNAseq data

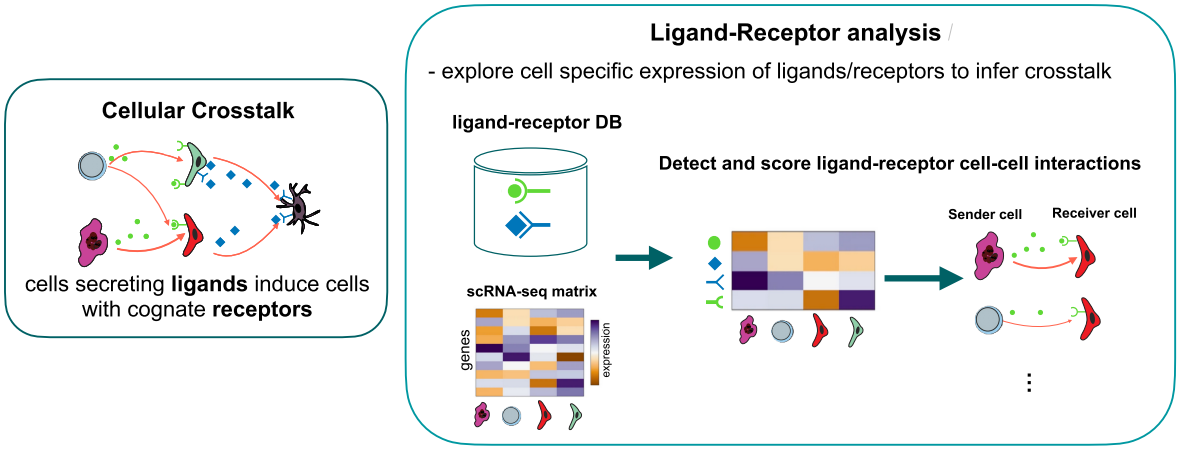


Figure 2.11: The overall CCC inference framework: How LR priors are combined with expression profiles for the identification of meaningful interactions, Adapted from (Armingol et al., 2020)

Given a scRNAseq dataset (2.4) with the respective clustering (2.5), to identify communication patterns across cell types, we perform CCC inference. Precisely, we can introduce the problem as defined below in the Definition 2.6.

Definition 2.6 (CCC inference problem)**.** *Let $X_{|g|,|sc|}$ be a single cell expression matrix as defined previously (see Definition 2.4) where cells have the clustering cl and ligand-receptor (LR) list $lr = (lr_1,\ldots,lr_{|lr|})$, as shown in Figure 2.11. A scoring function $f(s,g_{lig},r,g_{rec}) = w$ in which $s \in cl$ is the sender cell, $r \in cl$ is the receiver cell, g_{lig} represents a ligand gene, g_{rec} is a receptor gene and $w > 0$ is the weight of the interaction. An example of this score is the geometric mean between the average ligand/receptor expression within the source and target cell types. Given a set of all scored 5-tuples $IT = \{it_1,\ldots,it_n\}$ with $it = (s,g_{lig},r,g_{rec},w)$, the CCC inference problem consists in solving a function m which generates the subset $S \in IT$ containing the interactions of interest.*

In the current CCC inference literature, well-established tool examples are CellPhoneDB (Efremova et al., 2020), and CellChat (Jin et al., 2021). To provide a better understanding of how LR scores are computed and of the interaction selection criteria used in each tool, we describe the tools above.

CellPhoneDB

Developed by Efremova et al. (2020), CellPhoneDB (v1) not only provides a curated set of communicating gene pairs (e.g., LR pairs) but also introduces one of the pioneering works in CCC inference.

CellPhoneDB primarily uses scRNAseq and the LR database to compute the weights w and uses the function m to select meaningful interactions, as outlined in Definition 2.6. Here, the interaction scoring function $f(s, g_{lig}, r, g_{rec}) = w$ in which $s \in cl$ is the sender cell, $r \in cl$ is the receiver cell, g_{lig} represents a ligand gene, g_{rec} is a receptor gene is defined. CellPhoneDB, first computes the function 2.15 for all the pairs in the interaction list.

$$f(s, g_{lig}, r, g_{rec}) = w = \begin{cases} 0 & \text{if } \overline{X}_{g_{lig},sc=s} = 0 \text{ or } \overline{X}_{g_{rec},sc=r} = 0 \\ \frac{\overline{X}_{g_{lig},sc=s} + \overline{X}_{grec,sc=r}}{2} & otherwise, \end{cases} \tag{2.15}$$

where $\overline{X}$ represents the mean of the combined gene expression. Alternatively, the computed interaction scores can be represented in a matrix format $\mathbf{W}_{(g_{lig},g_{rec}),(s,r)}$ where rows represent the LR interactions and columns represent all possible cell pairs. The next step in CellPhoneDB is the interaction selection step. This step involves defining a function m that generates a subset of the interactions in $\mathbf{W}$.

Let $\mathcal{W} = (\mathcal{W}^1, \ldots, \mathcal{W}^p)$ be the set of the interaction matrices obtained after performing a permutation of cells and p the number of samples determined by the user and $\mathbf{W}$. Note that, here, the function m implements a permutation test and $\mathcal{W}$ represents a null distribution where the p-value is computed via:

$$p - value(s, g_{lig}, r, g_{rec}) = \frac{\#\text{ of } p \text{ where } \mathcal{W} \geq \mathbf{W}}{|\mathcal{W}|}. \tag{2.16}$$

A significant interaction (e.g., low p-value) indicates that the observed interaction is unlikely to occur randomly. Finally, m is a selection function that uses the computed p-value.

$$m(s, g_{lig}, r, g_{rec}) = \begin{cases} f(s, g_{lig}, r, g_{rec}) & \text{if } p - value(s, g_{lig}, r, g_{rec}) \leq 0.05 \\ 0 & otherwise. \end{cases} \tag{2.17}$$

CellPhoneDB implementation presents additional quality control steps. First, interactions are considered only if a gene is found in more than $x\%$ of the cells in the given cluster cl. Second, specific LR pairs involve a complex of receptors. CellPhoneDB uses the minimum expression value across genes in a gene complex. Details regarding these aspects are available in Efremova et al. (2020).

CellChat

Another state-of-the-art tool for the CCC inference is CellChat (Jin et al., 2021). In contrast to CellPhoneDB, CellChat relies on a set of interactions categorized according to known signaling pathways from a specific database. Before computing the interaction score w, CellChat considers the differentially expressed signaling genes, using a robust mean expression based on distribution quantiles. The Wilcoxon rank-sum test is then used to identify differentially expressed genes.

To combine the LR score per cell pair, the following function $f(s, g_{lig}, r, g_{rec}) = P^k_{i,j}$ is computed using the law of mass action. As a distinguishing feature, CellChat assumes that ligands and receptors can each have a set of subunits, denoted lm and rm, respectively. To this end, CellChat computes the ligand score L_i as:

$$L_i = \sqrt[|lm|]{L_{i,1} \cdots L_{i,|lm|}}. \tag{2.18}$$

Note that L_i represents the geometric mean; therefore, the zero expression of any subunit leads to an inactive ligand. For computing receptor scores, CellChat considers the receptors that excite or inhibit the cell interaction, which are denoted RA_j and RI_j, respectively, such that $RA_j + RI_j = |rm|$. This additional information is used to scale the geometric mean. Altogether, R_j is computed as:

$$R_j = \sqrt[|rm|]{R_{j,1} \cdots R_{j,|rm|}} \cdot \frac{1 + RA_j}{1 + RI_j}. \tag{2.19}$$

Now, by using the ligand/receptor scores as described above, the interaction score (e.g, communication probability) is computed as:

$$P^k_{i,j} = \frac{L_i R_j}{K_h + L_i R_j} \times \left(1 + \frac{AG_i}{K_h + AG_i}\right) \cdot \left(1 + \frac{AG_j}{K_h + AG_j}\right) \times \frac{K_h}{K_h + AN_i} \cdot \frac{K_h}{K_h + AN_j} \times \frac{n_i n_j}{|sc|^2}, \tag{2.20}$$

where K_h is a hill saturation constant, which ranges from 0 to 1. By default, CellChat adopts a half-saturation process ($K_h = 0.5$). Another additional feature adopted by CellChat is the use of agonist (AG) and antagonist (AN) ligands, which activate and block receptors, respectively. Finally, n_i/n_j denotes the number of cells that are labeled as $sc = s/sc = r$, respectively.

After computing all the $P^k_{i,j}$, a permutation test is used to select the significant interactions. This process is analogous to what is employed by CellPhoneDB (Efremova et al., 2020).

$$p - value(P^k_{ij}) = \frac{\#\text{ of } m \text{ where } \mathcal{P}^m_{ij} \leq P_{ij},\ m = 1, 2, \ldots, M}{|\text{no. of permutations (M)}|} \tag{2.21}$$

Notably, $\mathcal{P}^m_{ij}$ is the communication probability obtained after performing a cell type label permutation.

2.5 Discussion

In this chapter, we covered the essential concepts used throughout the thesis. As mentioned before, graph-related methods help uncover interaction patterns in various contexts (e.g., social networks, transportation networks, biological networks, and others). Even though there is rapid development in graph theory and related methods, complex systems are often represented as directed and/or signed networks, which yield strategies that account for the non-symmetric and non-linear nature of the data. The cell-cell communication network is a biological example of such complex data.

As mentioned above, measuring CCC at the single-cell level using proteins is a difficult task. To overcome this limitation, scRNAseq has been shown to be a useful proxy for inferring CCC. scRNAseq-derived CCC inference has improved our understanding of how cells communicate (Armingol et al., 2024). Although scRNAseq generates a snapshot of the transcriptome at a single time point (i.e., each measurement of each cell is a unique observation in time, given that cells are dead after the sequencing process), CCC inference methods have been successfully developed and applied in recent years (Efremova et al., 2020; Hou et al., 2020; Jin et al., 2021). A recurring issue with state-of-the-art LR inference methods (e.g., CellPhoneDB) is that their results are hard to interpret. The output of these tools is often a ranked list of hundreds of LR and cell pairs for a given scRNAseq dataset.

Because interpreting such ranked lists is challenging and non-intuitive, tools are needed to simplify these results, e.g., to prioritize cells and/or genes. Moreover, most LR inference methods focus on analyzing a single scRNAseq experiment for a single phenotype. They are unable to characterize changes in cellular crosstalk between pairs of phenotypes, e.g., disease versus homeostasis. Additionally, when working with patient data, there are only a few CCC tools that allow comparison of sample-specific CCC programs.

In summary, we have introduced the computational background and challenges related to the current thesis. In the computational aspect, we introduce node importance measures in graphs and discuss their relevance to real-world applications. Furthermore, we also cover concepts related to the graph comparison problem, namely, graph matching and graph optimal transport. In the CCC context, we present the motivation, data acquisition, and state-of-the-art methods for the CCC inference problem. In the upcoming chapters, we demonstrate how cellular communication networks can be compared and used to translate data-driven biological insights into improved medical treatments for complex diseases.

CHAPTER 3

Comparing pairs of cell-cell communication graphs

In Chapter 2, we introduce how graph comparison is a complex yet essential task. Comparing graphs can identify and rank changes in networked models, thereby expanding our understanding of how complex systems evolve. In computational biology, cell-cell interactions (CCI) are measured in several ways, such as cell-cell communication (CCC) or cell-cell co-localization.

In recent years, scRNAseq has enabled the inference of CCC from ligand–receptor (LR) gene expression profiles (Cesaro et al., 2025). Although CCC can occur in many ways, LR interactions are the primary type. CCC inference tools [e.g., CellPhoneDB (Efremova et al., 2020), NATMI (Hou et al., 2020), and CellChat (Jin et al., 2021)] couple a scRNAseq experiment and a known LR interaction database and output a ranked and scored list of hundreds of ligand-receptor and cell pairs. Scores generated by each CCC inference tool provide distinct readouts of how CCCs are observed in the environment, adding another layer of complexity to interpreting results.

Interpreting such ranked lists is challenging and sometimes unclear, necessitating tools to simplify these results, e.g., by prioritizing cell types and gene pairs from LR inference tools' results. To tackle this, CCC outputs can be represented and analyzed as graphs as shown in Figure 3.1. Furthermore, most LR inference methods focus on analyzing a single phenotype in scRNAseq experiments. They are unable to characterize changes in cellular crosstalk between pairs of phenotypes, e.g., disease versus control.

Differential CCC analysis paves the way for revealing key interactions and features across complex disease stages (e.g., healthy and diseased). By prioritizing key interaction patterns, a more effective biological validation is possible. Although such comparisons are highly desirable, there are only a few methods for comparing scRNAseq-derived CCCs. In the following chapter, we illustrate how we can compare graphs in the context of cell-cell communication, with a primary focus on pairwise comparisons. Next, we describe our method CrossTalkeR (Nagai et al., 2021), which enables the selection and identification of network differences in CCC graphs. At the end, two applications of CrossTalkeR for fibrotic diseases are shown.

3.1 Motivation and Related Work

CCC graphs have been used to augment knowledge of the events that trigger signalling pathways within a tissue. Among the challenges of using graphs to represent scRNAseq-derived CCC graphs, three main obstacles stand out for their complexity. First, the behaviour given by the direction of ligand-receptor pairs in the CCC imposes an asymmetric representation when constructing the CCC

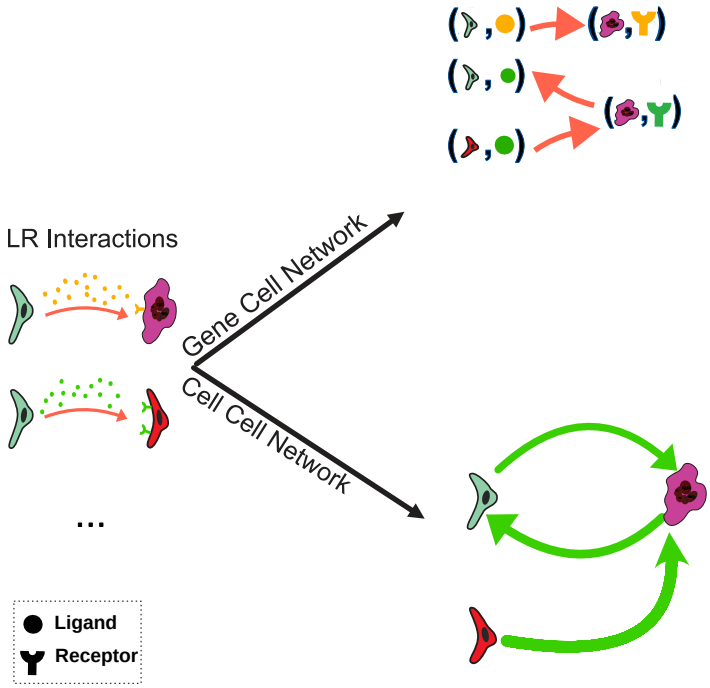


Figure 3.1: From LR interactions to Graphs: On the left, a schematic represents the output of CCC inference tools. On the right side, the two types of networks one can use to study further those interactions.

graph. Second, interactions at the cell-pair level can be measured in distinct ways, for example, the proportion of significant LR interactions and the total expression of LR interactions. Third, CCC inference methods often focus on meaningful LR interactions, so only a few methods are designed to provide condition-based comparative analysis.

Definition 3.1 (CCC graph)**.** *Let a cell-cell communication graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$*,* $\mathcal{G}$ *is a directed weighted graph where* $\mathcal{V}$ *is set of cell types,* $\mathcal{E}$ *is the set of cell-cell interactions and* $w : E \rightarrow \mathbb{R}^+$ *is a function that weights every interaction* $e \in \mathcal{E}$

By design, some scRNAseq-based CCC tools enable the construction of CCC graphs and the performance of differential CCC analysis. Among the large number of tools for CCC inference (Cesaro et al., 2025), NATMI [Hou et al. (2020)] and CellChat [Jin et al. (2021)] stand out as pioneering works in the field that use graphs to represent CCC outputs. Therefore, we revisit these methods in the following subsections.

3.1.1 CellChat

After selecting relevant LR interactions (see section 2.4), CellChat (Jin et al., 2021) results in a tensor $\mathcal{P}_{|cl|\times|cl|\times|z|}$ where $|cl|$ is the total number of cell types and z represents the number of signaling pathways/LR interactions. Each $\mathcal{P}_{|cl|\times|cl|\times|z_i|}$ is a CCC graph (as defined in Def. 3.1), where w maps each cell-cell interaction to the total number of selected LR pairs. To explore further topological properties of such a graph, the authors use the following centrality measures: (1) in/out-degree, (2)

flow-betweenness, and (3) information centrality. For comparing two distinct graphs, CellChat (Jin et al., 2021) presents two metrics: functional similarity (S) and structural similarity (D). Let G and G' be CCC graphs, and S is computed using the Jaccard similarity of the edges defined as:

$$S = \frac{E_G \cap E_{G'}}{(E_G \cup E_{G'}) - (E_G \cap E_{G'})}, \tag{3.1}$$

where E_G represents the edges of the CCC graph G, the same rationale applies to $E_{G'}$. High values of S indicate that the two graphs exhibit redundant functional roles. On the other hand, structural similarity (Schieber et al., 2017) treats graph topologies as probability distributions for graph comparison.

A key feature of CellChat is the use of additional annotations for ligand-receptor pairs, specifically indicating whether an LR pair triggers a signalling event (i.e., agonist) or blocks it (i.e., antagonist). This feature is implemented by taking into account the LR interaction annotation from KEGG Pathways database (Kanehisa, 2002). At last, CellChat (Jin et al., 2021) has been used to analyze cell-cell communication using scRNA-seq data, providing an easy-to-use toolkit and a well-defined framework.

3.1.2 NATMI

In contrast to the methods described in Section 2.4 in Chapter 2, NATMI (Hou et al., 2020) employs a graph-based approach to infer cell-cell communication. To this end, the LR scores are computed using the function $f(s, g_{lig}, t, g_{rec})$ where s is the source cell type, g_{lig} represents the gene expression value of the ligand lig in the cell type s, t is the target cell and g_{rec} represents the gene expression value of the receptor rec in the cell type t. Using these parameters, f can be computed using three distinct forms using the gene expression X.

- Geometric Mean: $f(s, g_{lig}, t, g_{rec}) = \overline{X}_{g_{lig}, sc=s} \times \overline{X}_{g_{rec}, sc=t}$
- Specificity: $f(s, g_{lig}, t, g_{rec}) = \frac{\overline{X}_{g_{lig}, sc=s}}{\overline{X}_{g_{lig}, s' \in sc}} \times \frac{\overline{X}_{g_{rec}, sc=t}}{\overline{X}_{g_{rec}, t' \in sc}}$
- Total Sum: $f(s, g_{lig}, t, g_{rec}) = \sum X_{g_{lig}, sc=s} \times \sum X_{g_{rec}, sc=t}$

Once the scores are computed using f, NATMI (Hou et al., 2020) applies a thresholding approach to select meaningful interactions based on the criteria above. For example, the filter selects interactions with a total count per million of at least 10, a specificity of 0.1, and a total count of 100. Formally, let et, st, pt be the thresholds adopted for the expression levels, specificity levels, and proportion levels, respectively. The selection function m is defined by:

$$m(s, g_{lig}, r, g_{rec}) = \begin{cases} f(s, g_{lig}, r, g_{rec}) & \text{Geometric Mean} >= et; \text{Specificity} >= st \text{ and Total Sum} >= tt \\ 0 & otherwise. \end{cases} \tag{3.2}$$

Next, the interactions obtained after computing the function m are summarized in a graph $G = (V, E)$ where nodes are cell types, and edges are placed according to the proportion of the cell pair $((s, t))$ interactions found in the set. Moreover, each edge in the graph presents the respective measurements

(i.e., Geometric Mean, Specificity, and Total Sum) as attributes. NATMI compares graphs using the so-called "delta network". Let G and G' be cell communication graphs, and G^{Δ} be the delta network:

$$E_{G^{\Delta}} = E_G - E_{G'}. \tag{3.3}$$

NATMI is the first CCC inference tool to distinguish between highly expressed interactions and numerically cell-pair-specific interactions. Furthermore, NATMI (Hou et al., 2020) provides a curated LR database called Connectome2020. NATMI introduces a selection function m that accounts for both the intensity and the specificity of the interaction. Moreover, NATMI offers better running time than state-of-the-art tools, such as CellPhoneDB. The scalability assay results are found in the supplementary notes from Hou et al. (2020). Finally, NATMI provides an LR database with literature-supported LR pairs.

3.1.3 Limitations of previous work

As introduced above, substantial work has been produced using graphs for CCC inference. In CellChat, node importance measures are used to rank cells, and an additional LR-pair feature is used for tailored CCC inference. On the other hand, node importance measures are not computed at the gene level, and CellChat does not implement node importance measures that enable comparisons between pairs of graphs. NATMI is a graph-based toolkit for the CCC inference. The tool focuses on inferring interactions using thresholding of different LR measurements (i.e., expression, sensitivity, and abundance). While NATMI is a pioneering tool for contrasting LR interaction features (e.g., intensity, quantity, and specificity), it does not provide methods for prioritizing condition-specific cell types or communication genes.

Despite NATMI and CellChat efforts, CCC interactions are complex and combinatorial, often requiring comparisons of graphs from different conditions to identify key changes in CCC. This calls for methods that perform both prioritization of important players (genes and/or cells) in CCC graphs and approaches that allow for comparing pairs of graphs, for example, a disease CCC graph with a control CCC graph. To fill this gap and provide a tailored framework for comparing CCC graphs, we introduce CrossTalkeR. CrossTalkeR leverages node importance measures to provide a comparative analysis in CCC graphs.

3.2 CrossTalkeR

Aiming to provide an easy-to-interpret prioritization of CCC inference results and to tackle the limitations described above, CrossTalkeR summarizes predictions from a CCC inference tool using a two-level graph approach: cell and cell/gene graphs. First, it uses specific network topology measures to rank cell types, cell receptors, and cell-ligand pairs. Second, CrossTalkeR enables researchers to perform a phenotype-based comparative LR analysis. CrossTalkeR is a complementary approach to current CCC inference tools. It is the only approach that explores LR networks at both the cell and cell-gene levels, comparing these networks across pairs of phenotypes using node importance

measures.

CrossTalkeR was developed in 2021 and has been maintained and developed as R and Python package. It presents the results either graphically or in tabular form, summarizes them in an HTML/PDF report, and stores all results in an R container to facilitate data sharing and reproducibility. In the next sections, we first describe how our method, CrossTalkeR, works. First, we describe how CrossTalkeR uses CCC inference output. Second, we will delineate the use of graph node-importance measures for performing network comparisons.

3.2.1 Network Construction

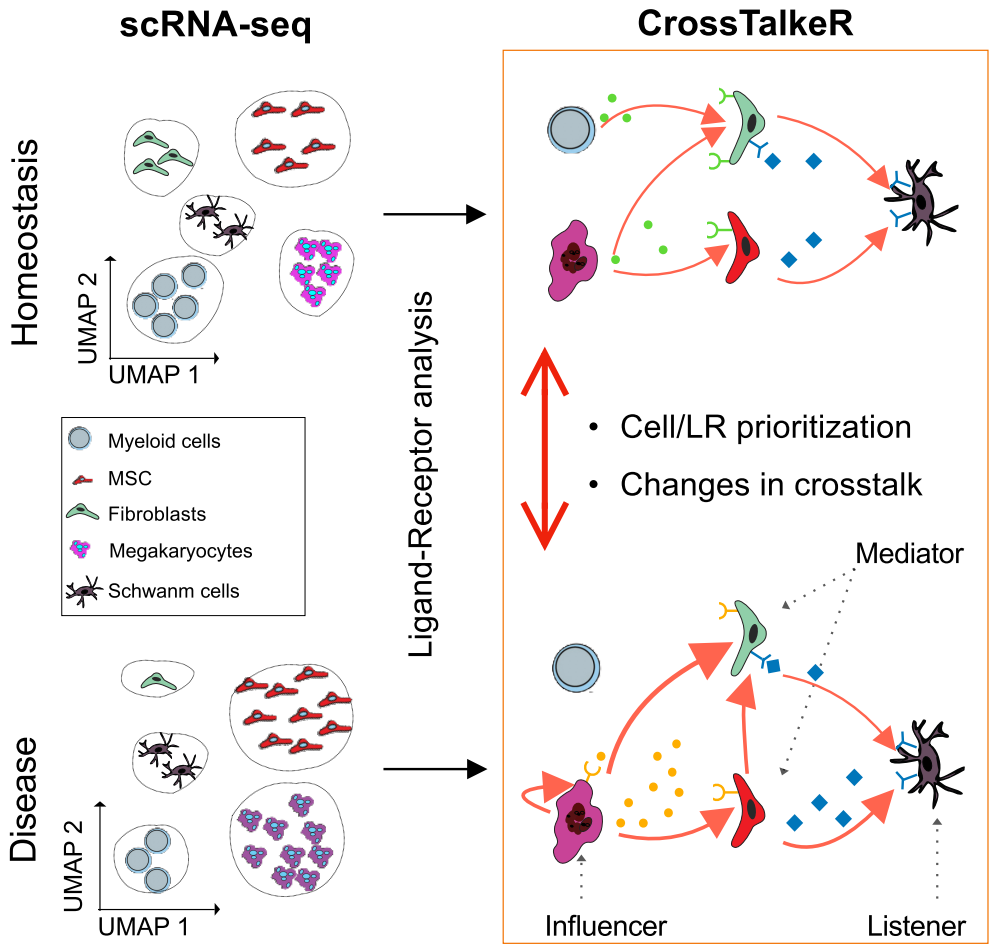


Figure 3.2: CrossTalkeR scheme: CrossTalkeR takes as input the CCC inference for each condition within an integrated scRNAseq dataset. Using the interaction list, the cell-level and gene-level graphs are built. Node importance measures (comparative and single-condition) are used for cell/gene prioritization and for pairwise condition comparisons.

Input Data

The input to CrossTalkeR consists of a list of 5-tuples $X = \{x_1, ..., x_n\}$ comprising n LR interactions predicted by an LR inference tool (e.g., CellPhoneDB) from a scRNAseq dataset. Here each 5-tuple

Table 3.1: CrossTalkeR input example obtained from the control samples from diseased bone marrow samples

Source	Target	Ligand	Receptor	MeanLR
MSC	Megakaryocyte	B2M	CD3D	2.89
Myeloid	MSC	S100A8	ITGB2	2.84
...	...	...	...	...
MSC	MSC	LGALS3	ENG	0.32

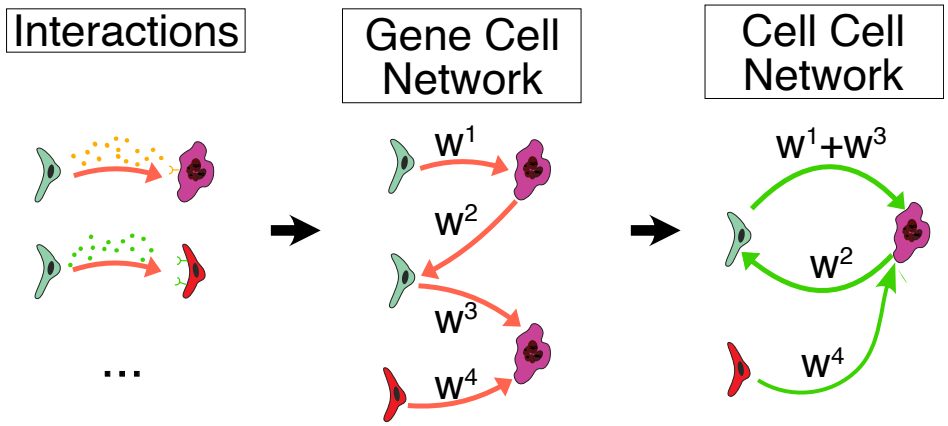


Figure 3.3: From LR inference results to graphs: Illustration of how CCN and GCN are built by CrossTalkeR

is of the form $x = (s, g_{lig}, t, g_{rec}, w)$, where $s \in \{1,..,k\}$ is the source cell, $t \in \{1,..,k\}$ is the target cell, g_{lig} represents a ligand gene, g_{rec} is a receptor gene and $w > 0$ is the weight of the interaction. The cell types (s and t) are defined as belonging to one of the k clusters detected in a given scRNA-seq dataset. The weight w of the interaction indicates the strength in the ligand-receptor signal, e.g., as estimated by the mean expression of the ligand and receptor (MeanLR) in CellPhoneDB. Note that interpretation of CrossTalkeR results will depend on the CCC inference tool's measurement (e.g., specificity and intensity). A sample of the input is represented in Table 3.1.

Generating the networks

As an initial step into the CrossTalkeR framework, the input table is used to construct networks at two different resolutions (Figure 3.3): the cell-gene interaction network (CGN) and the cell-cell interaction network (CCN). Formally, CGN and CCN are generated using the Definition 3.2 and Definition 3.3, respectively.

Definition 3.2 (Cell-Gene Network-CGN)**.**

The vertices V_{CGI} of the Cell-Gene interaction network consist of all possible pairs of ligands with sender cells (s, g_{lig}), and receptors with receiver cells (r, g_{rec}). Each tuple $(s, g_{lig}, r, g_{rec}, w) \in X$ now gives rise to one directed edge in a weighted graph with vertices V_{CGI}. Specifically, the entries of weighted adjacency matrix $\mathbf{W_{CGI}} = \{w_{i,j}\}^{n \times n}$ are defined as $w_{(g_{lig},s),(g_{rec},r)} = w$.

Definition 3.3 (Cell-Cell Network-CCN)**.**

We obtain a CCI network by transforming the list of 5*-tuples X into a list of triplets. Each of these triplets is of the form* $t = (s, r, w_{s,r})$*, where s represents the sender cell type, r the receiver cell type, and* $w_{s,r}$ *represents the weights of the interactions from cell type s to cell type r. Let* $W_{s,r}$ *be the set of all weights, where s is the sender and r is the receiver cell,*

$$W_{s,r} = \{w | (s', g'_{lig}, r', g'_{rec}, w) \in X, s' = s, r' = r\}. \tag{3.4}$$

Based on these sets, we can define a new weighted directed graph between cells. The vertex set of this graph V_{CCI} *consists of the set of all cell types* $\{1, .., k\}$*, and we denote its adjacency matrix as* $\mathbf{W}' = \{w'_{s,r}\}^{k \times k}$*, with entries defined as:*

$$w'_{s,r} = \sum_{w \in W_{s,r}} w. \tag{3.5}$$

Alternatively, the adjacency matrix can be written in terms of the LR set cardinality:

$$w'_{s,r} = |w \in W_{s,r}|. \tag{3.6}$$

3.2.2 Differential Network Analysis

Once both CGN and CCN are assembled, we can perform the differential analysis. To this end, we assemble our differential network as defined in Definition 3.4.

Definition 3.4 (Building Differential Network)**.** *Given two CGN networks with adjacency matrices* $\mathbf{W}^{\mathbf{exp}}$ *(experimental condition) and* $\mathbf{W}^{\mathbf{cont}}$ *(control), defined over the same vertex space V, we can obtain a differential network as* $\mathbf{W}^{\mathbf{exp}} - \mathbf{W}^{\mathbf{cont}}$*. Here, positive values indicate interactions with higher weights in the experiment, and negative values indicate interactions with higher weights in the control. The same follows for the CCN,*

Notably, it is essential to ensure that interaction predictions in the experiment $\mathbf{W}^{\mathbf{exp}}$ and control $\mathbf{W}^{\mathbf{cont}}$ conditions are performed on integrated, normalized single-cell RNA-seq data to ensure comparable weights and cell/gene spaces. An example of a valid integration and normalization strategy using Seurat is found in our CrossTalkeR tutorial (based on the PMF data).

Implementation of Network Topological measures

A key feature of CrossTalkeR is the use of node importance measures to prioritize cells/genes in scRNAseq CCC data. To this end, for each single biological condition (single phenotype) the CCN or CGN are used for computing the following measures (1) indegree d_{in} (listener score); (2) outdegree d_{out} (influencer score); (3) PageRank (importance score) p and (4) betweenness centrality b (mediator score). CCN and CGN, in their natural representation, are directed and weighted; therefore, we need to define node importance measures suitable for such a data representation. For PageRank and betweenness, we refer to the definition presented in Chapter 2.2.

Definition 3.5 (In/out Degree)**.** *Let* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ *be a directed weighted graph as CCN/CGN with* $\mathcal{V}$ *representing the set of nodes (cells/genes) and* $\mathcal{E}$ *representing the interactions placed using the*

CCC inference selected LR pairs. $\mathcal{W}$ *represents the weighted adjacency matrix. The in/out-degree vector can be computed as follows:*

$$d_{in} = \mathcal{W}^T \mathbf{1}, \tag{3.7}$$

where $\mathbf{1}$ *is a matrix of ones. For outdegree, we can compute the vector by* $d_{out} = \mathcal{W}\mathbf{1}$.

By performing CCC inference on a normalized, integrated data CCI network for each condition, one observes two interesting properties. First, once the cell type annotation is universal across conditions, the cell type set (i.e., nodes from the CCN) is similar. Second, once the dataset is normalized, the LR interaction values are sampled from the same gene-expression distribution. Altogether, these properties allow us to obtain both CCN and CGN, with features in the same range and node sets with a high intersection. When comparing two phenotypes (experiment and control), both CGN and CCN are signed and directed, and there is no standard procedure for analyzing such signed networks in the same way as networks with non-negative edge weights. Precisely, let us consider two biological conditions *control* (*ctr*) and *experiment* (*exp*); for each condition, CrossTalkeR generates a CGN and a CCN graph. As the next step, using both graphs, CrossTalkeR computes the following differential node importance measures.

$$\Delta_{in} = d_{in}(exp) - d_{in}(ctr) \tag{3.8}$$

$$\Delta_{out} = d_{out}(exp) - d_{out}(ctr) \tag{3.9}$$

$$\Delta_{betweenness} = b(exp) - b(ctr) \tag{3.10}$$

Pagerank evaluates the importance of nodes in a graph by exploring its connections, and is equal to the stationary probability distribution of a certain diffusion process on the network (Gleich, 2015). When applied to either the experimental condition or the control network, the PageRank of a node indicates the probability that the node will be visited by a diffusion process with random restarts (Page et al., 1998; Gleich, 2015) and can be used as a proxy for node importance in both CCN and CGN networks. For a given cell c and state $s \in \{control, exp\}$, PageRank returns a value $\text{Pagerank}_s(c)$, which is proportional to the probability of the diffusion process being at a cell c for a given state s.

$$P(c|s = exp) \approx \frac{\text{Pagerank}_s(c) + \alpha}{\sum_{c\prime}(\text{Pagerank}_s(c\prime) + \alpha)}, \tag{3.11}$$

where α is a regularizing constant. This regularization is necessary because the networks do not necessarily share the same vertex set; i.e., a cell or a cell/gene pair is not present in a phenotype. $P(c|s = control)$ can be estimated accordingly. Since PageRank can also be interpreted as a probability distribution (Page et al., 1998), we used log-odds to measure the difference in node importance between the PageRank conditions *exp* and *control*.

$$logodds(c) = \log\left(\frac{P(c|s = exp)}{P(c|s = control)}\right). \tag{3.12}$$

Positive posterior log-odds indicate a higher probability of accessing a node in the experimental

condition network. In comparison, negative values indicate a higher probability of accessing a node in the control network. We can therefore identify nodes whose importance changes between two conditions. To further increase the biological readouts from CrossTalkeR, we implement two additional features for exploratory data analysis. The principal component analysis of the node importance measures aims to provide a visualization of the communication nodes (cells or genes). In a complementary manner, the KEGG enrichment provides the pathways associated with genes found in the top and bottom 100 interactions per node importance measure. We aim to provide a clue about the process triggered by each measure.

Principal Component Analysis (PCA) of network properties

CrossTalkeR employs PCA to combine distinct network measures for CGN/CCN. This analysis provides a visual way to identify features that explain the network's distinct (and independent) properties. Formally, let $\mathbf{M}_{v\times ni}$ be a matrix with v rows and ni columns, v and ni represent the nodes (cells or genes) and the normalized node importance measures, respectively. $\mathbf{M}_{v\times ni}$ is used for the PCA, which computes a SVD of the centered $\mathbf{M}_{v\times ni}$ matrix by $\tilde{\mathbf{M}}_{v\times ni} = U\Sigma V^T$.We use the two components with high variance to embed the nodes in a 2D space. Note that, for CCN, we only indicate nodes whose values stand out, i.e., whose PC1 or PC2 values are outside the 95% confidence interval assuming the PCs follow a Gaussian distribution.

KEGG Annotation form Top Gene Cell pairs

To provide a functional annotation, the top n (n = 100, as default) scored receptors/ligands for each topological measure (influencer/listener/mediator/PageRank) are used as input to the enrichKEGG function from the clusterProfiler R package (Yu, 2024). In short, we use the hypergeometric test to check if the top-ranked ligand/receptors for each topological measure associated with a particular KEGG pathway are higher than expected by chance. For the contingency table, we consider only genes annotated as ligands and receptors in CellPhoneDB (Efremova et al., 2020). We then display enriched KEGG pathways as a heatmap. Additionally, we provide a list of pathways used to filter meaningful signaling pathways from the database.

3.2.3 Statistical test for comparing CCC

As mentioned before, measurements in CCC can capture distinct features in the interaction. One open question still is the trade-off between the proportion and the intensity of the CCC programs. To address this gap, we implemented two statistical tests to compare changes in CCC with respect to proportion and expression, respectively.

Differential CCI proportion test: Given $c \in C$ where $C = \{\text{ctr}, \text{exp}\}$, we consider the list of interactions of each condition as in Definition 3.3. In other words, the list of triplets t_{ctr} indicates the interactions of the control samples. For comparing the proportions, $w_{(s,r)}$ represents the number of LR interactions of a respective sender (s) and receiver (r) pair. Next, for each CCI $(s, r) = (k, l)$, we can compare the proportions using Fisher's exact test using the following hypothesis test.

$$H_0 : w^{c=\text{ctr}}_{(s,r)=(k,l)} = w^{c=\text{exp}}_{(s,r)=(k,l)}$$
$$H_1 : w^{c=\text{ctr}}_{(s,r)=(k,l)} \neq w^{c=\text{exp}}_{(s,r)=(k,l)},$$

where $p^{c=\text{ctr}}_{(k,l)}$ denotes the proportion of the (k,l) regarding the control samples, the same notation follows for disease samples. For computing the test significance, we consider the following contingency table:

	c=control	c=disease
$w_{(s,r)=(k,l)}$	c_{11}	c_{12}
$w_{(s,r)\neq(k,l)}$	c_{21}	c_{22}

The test significance (p-value) can then be computed as:

$$\text{p-value} = \frac{(c_{11}+c_{12})!(c_{21}+c_{22})!(c_{11}+c_{21})!(c_{12}+c_{22})!}{c_{11}!c_{12}!c_{21}!c_{22}!n!}, \tag{3.13}$$

where $n = c_{11}+c_{12}+c_{21}+c_{22}$, and c_{11} indicates the amount of LR pairs within the (k,l) interaction from the control and c_{12} follows but from the disease, the interpretation follows for the remaining entries of the contingency matrix. As a result, a table containing the changes in proportion and the test significance for each (s,r) is generated.

Differential CCI expression test: Another way to compare CCI within paired conditions is by comparing the changes in expression levels of the cell-cell communication pairs. To this end, we consider the tuples defined in Definition 3.2. Let F define the list $F^{c=ctr}_{(s,r)=(k,l)}$ of all LR pairs related to the (k,l) cell-cell interaction from the ctr condition and $F^{c=exp}_{(s,r)=(k,l)}$ for the exp condition. Each entry of the list $F^{c=ctr}_{(s,r)=(k,l)}$ corresponds to an expression of a specific LR pair.

To compare the distributions across paired conditions, we can use the following hypothesis test as defined in the Mann-Whitney U test.

$$H_0 : F^{c=\tilde{ctr}}_{(s,r)=(k,l)} = F^{c=\tilde{exp}}_{(s,r)=(k,l)}$$
$$H_1 : F^{c=\tilde{ctr}}_{(s,r)=(k,l)} \neq F^{c=\tilde{exp}}_{(s,r)=(k,l)}.$$

To compute the significance of the test, all the entries of $F^{c=ctr}_{(s,r)=(k,l)}$ and $F^{c=exp}_{(s,r)=(k,l)}$ are ranked; the lowest value ranks 1, and the largest value ranks n_1+n_2. Note that, $n_1 = |F^{c=ctr}_{(s,r)=(k,l)}|$ and $n_2 = |F^{c=exp}_{(s,r)=(k,l)}|$. Using the ranked data, the sum of ranks from the ctr samples is $R^{c=ctr}_{(s,r)=(k,l)}$. As a next step, the U-statistic is computed by:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R^{c=ctr}_{(s,r)=(k,l)},$$
$$U_2 = n_1 n_2 - U_1,$$
$$U = \min(U_1, U_2).$$

For large samples, the U-statistic approximates the normal distribution $\mathcal{U} \approx \mathcal{N}(\mu_{\mathcal{U}}, \sigma^{\epsilon}_{\mathcal{U}})$. Finally, the p-value is obtained by:

$$\text{p-value} = 2 \times P(|Z| \geq |Z_{\text{obs}}|),$$

where,

$$Z_{obs} = \frac{U - \mu_U}{\sigma_U}.$$

As a result, a table containing the changes in expression and the test significance for each (s, r) is generated.

3.2.4 CrossTalkeR technical requirements

For performing CrossTalkeR, we consider that the scRNAseq is integrated and normalized. This can be achieved in practice via the algorithm NormalizeData from the Seurat package with the following parameters: normalization.method = "LogNormalize" and scale.factor = 10000 (Stuart et al., 2019).

3.2.5 Current state of CrossTalkeR and Next Features

Since its development, CrossTalkeR has been constantly maintained and improved. CrossTalkeR was initially developed as an open-source R package available at https://github.com/CostaLab/CrossTalkeR. Main improvements in CrossTalkeR include (1) Implementation of statistical tests; (2) Distinction of gene types; and (3) Implementation of the comparable Python version of the package. To this date (02 October 2025), CrossTalkeR has 63 citations, 9 forks, 47 stars, 4 branches and more than 500 commits.

3.3 Applications using CrossTalkeR in Fibrotic Diseases

Currently, CrossTalkeR is available as a R/python package. Since its release, CrossTalkeR has been used in several studies with a diverse range of diseases. In the applications where CrossTalkeR was useful, we will mainly focus on fibrotic diseases here. Thus, we introduce fibrosis, a biological process that occurs in a wide range of severe diseases. Next, we illustrate the use of CrossTalkeR by revisiting our published scRNA-seq data. At the end, we illustrate two success cases where CrossTalkeR played a pivotal role in translating computational findings into real-world, medicine-relevant data.

3.3.1 Fibrosis

Fibrosis is a typical biological process that occurs as part of several chronic diseases and is understood as a result of an unregulated tissue repair response to injury (Rieder et al., 2025). In short, fibrosis occurs through the cumulative deposition of connective tissue (e.g., collagen) after injury. An example of fibrosis progression is shown in Figure 3.4. The wound-healing process has four main steps that precede fibrosis. In homeostasis, bleeding is stopped, and the platelet-fibrin clot acts as a healing scaffold. This stage is followed by inflammation, during which immune cells remove cellular debris and coordinate tissue repair. The proliferation stage follows: new blood vessels form, nearby cells

migrate into the wound, epithelial cells regenerate, and fibroblasts produce collagen. In the last step of wound healing, collagen is organized into scar tissue, excess extracellular matrix (ECM) is cleared, and related cells are repaired by apoptosis, often leaving the wound healed without scars.

Fibrosis is a process in which the repairing-related cells remain in the tissue, yielding a fibrotic repair characterized by ECM production and tissue stiffening, resulting in a hypertrophic scar (Younesi et al., 2024). In a fibrotic process, several cell types contribute to each one of the four steps mentioned above. One particular manner in which cells orchestrate such complex processes involves ligand-receptor CCC. A classic example of this is the release of inflammatory ligands by aberrant cells, such as TGFB1 or TNF-α, during the inflammatory step (Deng et al., 2024). Even though initial knowledge of the different communication patterns within fibrosis exists, many mechanisms remain uncovered across several distinct diseases. Examples of diseases associated with fibrosis include chronic kidney disease (CKD), primary myelofibrosis of the bone marrow, inflammatory bowel disease, and others (Rieder et al., 2025).

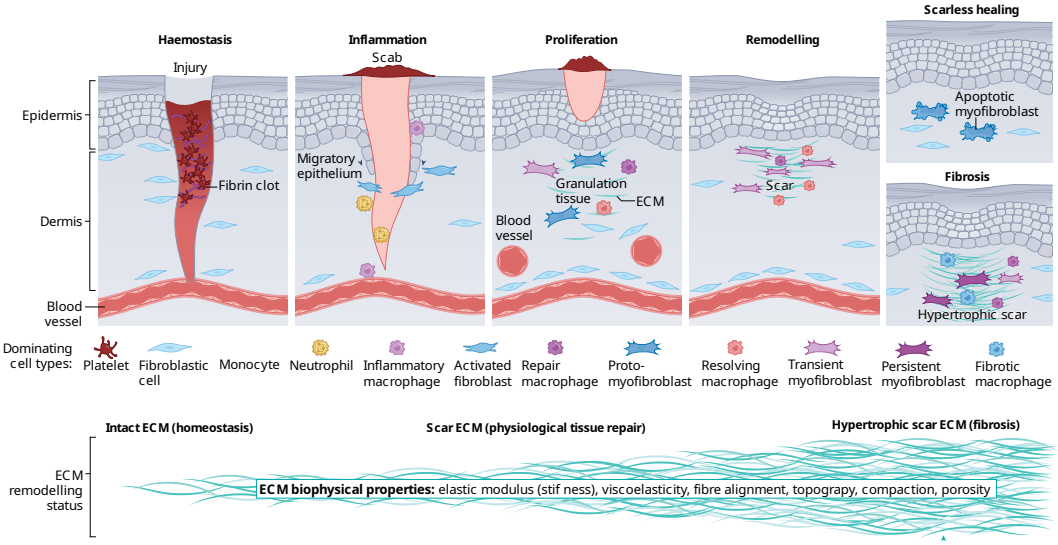


Figure 3.4: Fibrosis formation example in the context of wound healing. Adapted from Younesi et al. (2024)

We next present two case studies where CrossTalkeR was used. First, we show the results of CrossTalkeR in the context of bone marrow fibrosis. Second, we show how CrossTalkeR was used in the critical context of COVID-19, a global pandemic disease.

Bone Marrow Fibrosis

Within the diverse context of blood cancers, Myeloproliferative neoplasms (MPNs) are a heterogeneous group of cancers in which hematopoiesis (i.e., blood formation) is impaired, leading to abnormal production of myeloid cells. MPNs are divided into three diseases: (1) polycythemia vera, (2) essential thrombocythemia, and (3) primary myelofibrosis (PMF). PMF is the prototypical example of progressive BM fibrosis. One of the main characteristics of PMF is the excessive deposition of ECM, accompanied by progressive loss of hematopoiesis and splenomegaly due to extramedullary

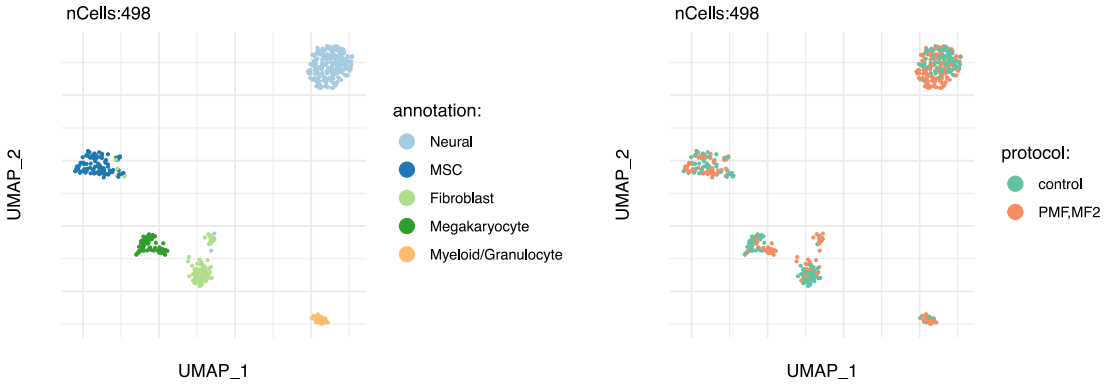


Figure 3.5: Points in the UMAP represent single cells. On the left side, every single cell is colored by the corresponding cell type. On the right, cells are colored by the disease stage of each sample. Adapted from Leimkühler et al. (2021)

hematopoiesis. Within the context of collaborative work with the Institute of Cell and Tumor Biology from the RWTH University Hospital, we performed scRNA sequencing using three biopsies (controls (n=2) and fibrosis (n=1)) from bone marrow samples (Leimkühler et al., 2021). In this manuscript, the biological assays, scRNAseq analysis, and project lead were performed by Dr. Nils Leimkuhler, Dr. Helene Gleitz, and Dr. Ronghui Li, and I contributed by performing the CCC inference and CCC analysis. Sequencing was performed to capture only non-hematopoietic cells, meaning that we do not target blood lineage cells. In total, we retrieved 498 single cells, and scRNAseq clustering was able to reveal five major cell types, shown in the Figure 3.5: (1) Mesenchymal Stromal Cells (MSCs), (2) Fibroblasts (Fib), and (3) Schwann cells (SC). For the markers used in the cell type annotation process, we refer to Leimkühler et al. (2021).

Using this rich data, we next interrogate what CCC programs occur in the disease. To this end, I performed CCC inference using CellPhoneDB (Efremova et al., 2020), and the differential network analysis CrossTalkeR. Figure 3.6 shows the CCN plots using a circle layout, with condition differences highlighted in the differential network. In the differential CCN, the cancer cell type (i.e., MK-Megakaryocytes) shows a high number of up-regulated autocrine and intercellular interactions. This results in a high PageRank log-odds for this cell population, clearly highlighting MK as the most essential signaling cell in this disease, influencing a wide range of bone marrow cells, including MSCs and Fibroblasts.

We also observe a significant reduction in interactions between MSCs and fibroblasts compared to homeostasis. This finding is in line with the restricted myofibroblastic phenotype that MSCs have been described to acquire. MSCs are usually strongly interconnected, moderating the quiescence and differentiation of hematopoietic cells, but lose this capacity as myelofibrosis progresses (Leimkühler et al., 2021; Schneider et al., 2017).

PCA analysis of the CGN network (Figure 3.7A) highlights matrix-related proteins (COL1A1, COL1A2), which are secreted by MSCs, and TGFB1 secreted by MK, to act as both influencers and mediators in myelofibrosis. Note that these findings recover the ECM behavior in the process of fibrosis as in Figure 3.4. These results follow known mechanisms associated with MF, e.g., crosstalk between MK (malignant cells) and MSCs via TGFb signalling, and the deposition on the matrix via

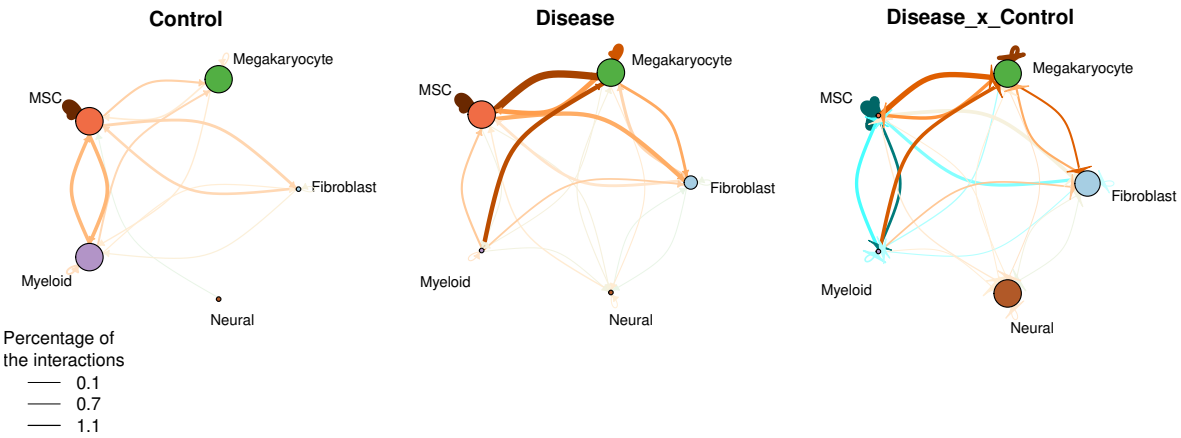


Figure 3.6: Cell-Cell Communication from BM fibrosis assembled by CrossTalkeR. The differential network provides a visual aid for analyzing the main differences in CCC. Nodes represent cell types; While in the single-condition graphs, edges are colored with aggregated gene expression, in the comparative graph (on the right), the edge colors represent the differences in gene expression. Edges colored in blue have a higher signal in the control, and the edges in orange have a higher signal in the disease. Edge thicknesses are set based on the number of LR pairs in the corresponding CCN. Node size is placed according to the PageRank vector; large nodes have greater importance for the CCC.

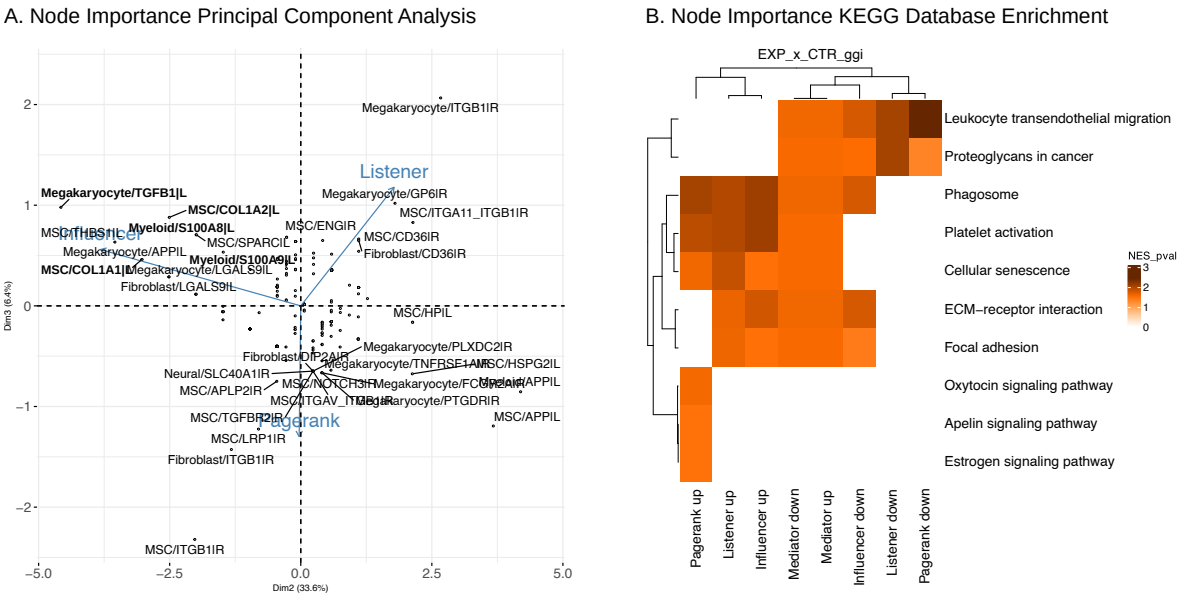


Figure 3.7: A-Visualization of the first two principal components obtained when using PCA over the node importance measures; elements in bold highlight BM fibrosis targets of interest. B-A heatmap containing the enrichment analysis using the KEGG database.

MSC cells (Gleitz et al., 2018).

KEGG pathway analysis (Figure 3.7B) supports these interpretations with up-regulated mediators and influencers mainly being enriched for Extracellular Matrix (ECM) deposition, while down-regulated influenced nodes are associated with leukocyte migration. LR analysis with CrossTalkeR also revealed that genes with increased importance are enriched for the KEGG pathway “platelet activation” (Figure 3.7B). These results recapitulate the finding that platelet-related mediators, such as

CXCL4, contribute to the persistent inflammatory environment in myelofibrosis (Gleitz et al., 2020). Altogether, our analysis recapitulates PMF biological aspects and comprehensively illustrates how CrossTalkeR can characterize changes in diverse cellular crosstalk across conditions from scRNAseq data.

Further investigation of TGFB signalling LR interactions in the dataset provided by CrossTalkeR also provides insight into the differential intercellular signaling present in the myelofibrotic bone marrow (Figure 3.8). The strong upregulation of TGFB1 between MK and MSCs is in line with other publications, showing canonical TGFB1-signaling as the most dominant pro-fibrotic element (Teodorescu et al., 2020). Thus, CrossTalkeR captures the multiplicity of signals present in biological networks at any biological condition.

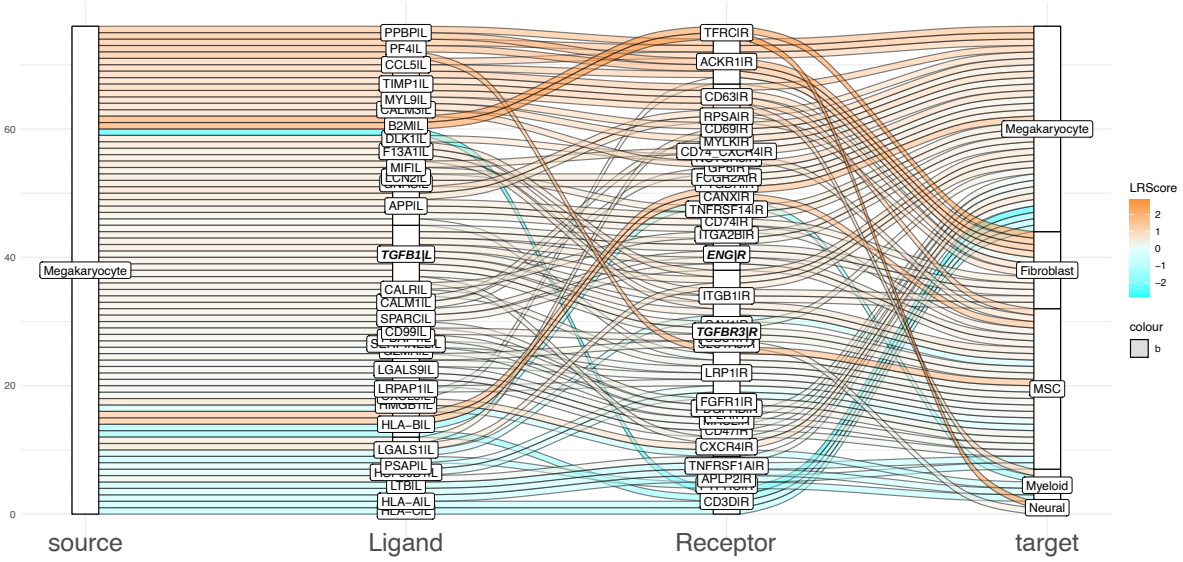


Figure 3.8: TGFB1 shows a high participation in the high-expression pairs. The Sankey plot shows the LR flow between cell types, and the diagram shows cells and receptors with higher LRScore values.

New insights into bone marrow fibrosis

The results presented above are consistent with prior knowledge of primary myelofibrosis in the bone marrow. Using CrossTalkeR, we identified the biological mechanisms underlying the pathology. Altogether, this indicates CrossTalkeR's ability to identify meaningful biological patterns. We next highlight results in which we were able to provide new knowledge into the PMF pathology.

While myeloid cells are among the leading players in PMF, the CCC related to myeloid cells remains unclear. CrossTalkeR results capture the alarmin heterocomplex S100A8/S100A9 related to Myeloid cells, as bolded in (Figure 3.7A). Sankey plots Figure 3.9 indicate MK and Fibroblast being central receptors of alarmin signaling.

Motivated by the literature (Razmkhah et al., 2023) and the CrossTalkeR results, Dr. Nils Leimkuhler and Dr. Helene Gleitz further examine the effects of the inhibition of the S100A8/S100A9 signaling process in the PMF context. To achieve this, they treated PMF mice with Tasquinimod. Tasquinimod binds to the S100A8/S100A9, blocking the interactions with its receptors Isaacs et al. (2006).

As shown in Figure 3.10A, there is a reduction in spleen size when PMF mice are treated with

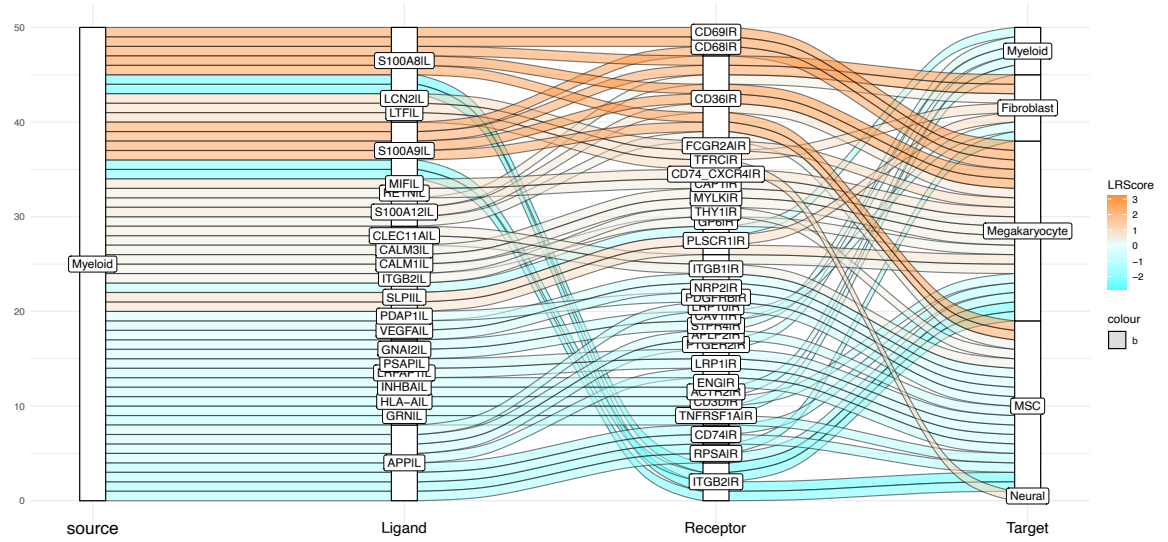


Figure 3.9: Myeloid cells top interactions S100A8/S100A9 are among the highest expressed LR pairs. Sankey plot implemented by CrossTalkeR for showing the LR flow between cell types. The diagram shows cells and receptors that have a higher LRScore.

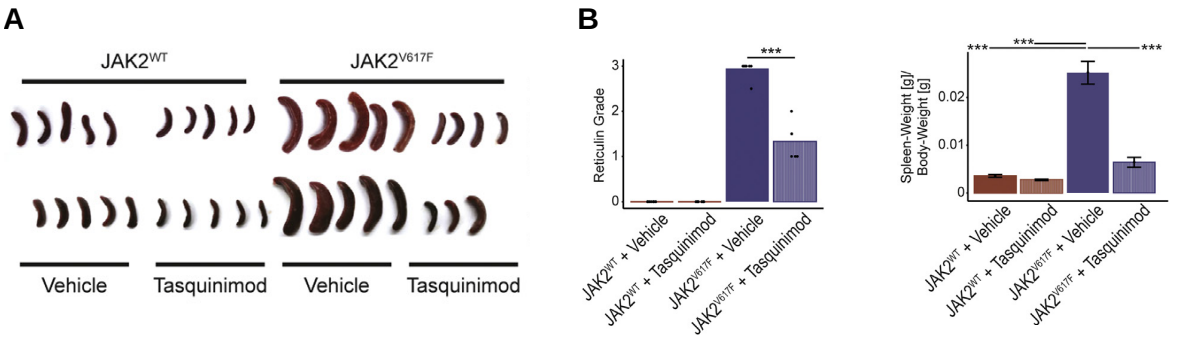


Figure 3.10: Revealing S100A8/S100A9 as fibrosis signatures: A. Mouse experiments, spleen samples showing the treatment effects in spleen size upon S100A8 inhibitor treatment, and B. Statistical comparison of spleen size and reticulin grade in the Tasquinimod-treated mouse samples.

Tasquinimod. Note that a typical phenotype of PMF is severe spleenomegaly. Those differences were also quantified as in the left side of Figure 3.10B. Another quantification is the amount of scarring in the tissue; therefore, they also quantify changes in the reticulin grade (i.e., amount of fibrosis) of the murine spleen. Altogether, the results established CrossTalkeR as a tool for revealing condition-specific differential communication patterns. Further research on the PMF-related findings from CrossTalkeR demonstrated the possible origin of the abnormal inflammatory process in the context of PMF (Gleitz et al., 2025). Furthermore, a clinical trial, HOVON HO172 (https://hovon.nl/en/trials/ho172), was conducted for further investigation into repurposing Tasquimod for PMF patients.

Revealling Kidney Fibrosis in SARS-COV-2 infected Kidney Organoids

Another use case for CrossTalkeR took place during the 2020 coronavirus disease pandemic (COVID-19). COVID is a flu-like respiratory syndrome that primarily affects the respiratory system but can also injure other organs. During the COVID-19 outbreak, kidney failure was frequently observed during and after the COVID-19 infection, but it remains elusive whether this is a direct effect of the virus. To better understand the biological process underlying the kidneys, Jansen et al. (2022)[1] performed scRNAseq in kidney organoids and COVID-19 patient kidney autopsies.

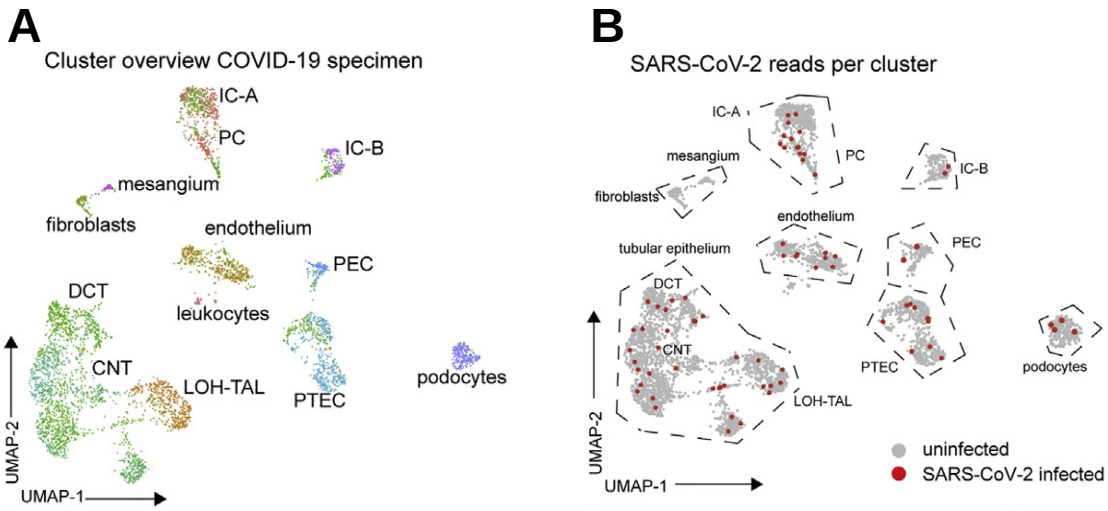


Figure 3.11: A. scRNAseq dataset UMAP from control and COVID samples presents the obtained cell populations. B. UMAP presenting the single cells in which viral reads were found.

In a collaborative effort with the Institute of Experimental Medicine and System Biology from the RWTH University Hospital, we combined targeted RNA sequencing and standards scRNAseq (see Figure 3.11), we were able to characterize distinct kidney cell populations and the potential viral reads per single cell, as shown in the Figure 3.11B. For the complex task of viral read identification, an additional sequencing protocol, TAPseq (Schraivogel et al., 2020), was used. Combining the design of custom adapters with a computational pipeline, TAPseq enables the identification of COVID-19 viral sequences in scRNAseq data. The details of the TAPseq analysis in the context of this manuscript can be found in (Jansen et al., 2022).

Using the valuable information from the TAPseq on infected cells, we were able to perform CCC on infected kidney cells. CrossTalkeR's node statistics indicate the importance of distinct CCC players when comparing COVID and control samples, see Figure 3.12A. Enrichment of the top 3 PageRank, highlighting ECM-related interactions, indicating the presence of a fibrotic process. Further analysis of PT cells to fibroblasts interactions revealed, in particular, fibroblast growth factor receptor (FGFR), EGFR, and platelet-derived growth factor receptor alpha (PDGFRa) signaling to be upregulated, see Figure 3.12B and C. To sum up, this study and CrossTalkeR were not only able to show the viral

[1] In this project led by Dr. Jitske Jansen, Dr. Katarina Reimer and MSc. James Nagai (myself) contributed as both the main bioinformatician of the project and to the CCC inference and analysis. Dr. Jansen and Dr. Reimer performed all the bench experiments, including sequencing, organoids, and autopsy samples.

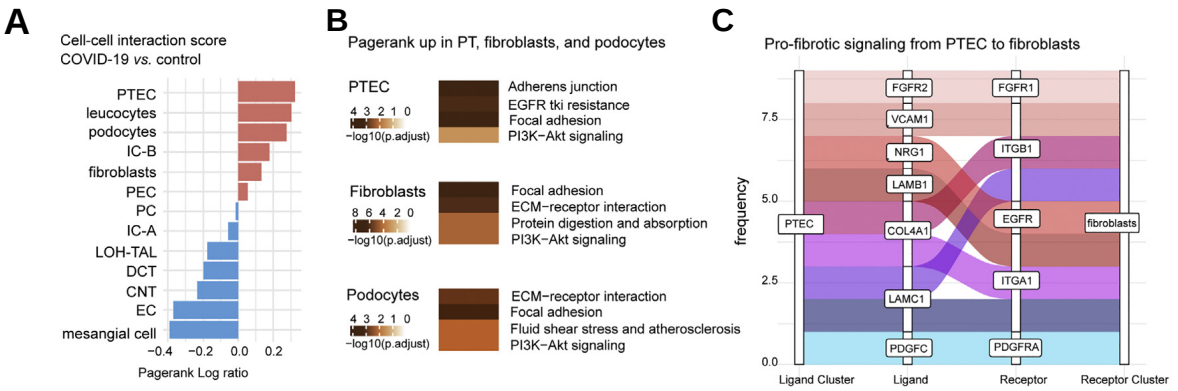


Figure 3.12: A. CCN node importance (Pagerank) indicates the most important cell populations in COVID. B. Pathway enrichment of main cell populations shows high association with ECM interactions. C. Sankey Plot showing the main LR interactions between PTEC and Fibroblasts.

infection effects on the immune cells but also revealed that PT cells are infected and promote fibrotic phenotype in a COVID-19 viral infection. Furthermore, the study also used kidney organoids (synthetic organs produced from pluripotent stem cells) to assess and reproduce COVID-19 infection in the kidneys. To provide tangible evidence of fibrosis in COVID patients, our study also includes data from more than 60 samples from infected patients and 57 controls. When comparing samples, we also observe increased fibrosis in infected patients. This corroborates CrossTalkeR's findings.

3.4 Discussion

As shown in this chapter, modeling and analyzing scRNA-seq-derived CCC are important tasks. Using graph methods, several tools yielded important observations at the cell-type or gene level. We also introduce state-of-the-art tools that use networks to highlight differences by applying filtering strategies such as NATMI (Hou et al., 2020). Moreover, tools in which the network topology was used but interactions were constrained to a restricted set of ligand-receptor interactions, as in CellChat (Jin et al., 2021). At the state of the art in graph approaches and to the best of our knowledge, CrossTalkeR is one of the pioneering tools for comparing pairwise conditions using node importance measures at multiple levels, including CCN and GCN levels. Finally, CrossTalkeR is an R package that has been maintained and improved since its first publication. CrossTalkeR allows both computer scientists and researchers in biology and medicine to analyze CCC inference data.

CrossTalkeR contributions can also be seen in several other manuscripts. For example, as illustrated by the aforementioned case studies, CrossTalkeR played a key role in identifying the target S100A8/S100A9. By further investigating this target, we demonstrated that Tasquinimod, an S100A8/S100A9 inhibitor, is a potential treatment for this specific type of blood cancer. Furthermore, the results of this study enabled the design of a clinical trial for PMF patients using Tasquinimod. In the COVID pandemic, CrossTalkeR played an essential role in revealing fibrotic patterns in the Kidneys of COVID patients. Our study not only detected the virus in renal tissue but also showed that COVID patients undergo a fibrotic process. To sum up, CrossTalkeR is an established tool that addresses the problem of comparative cell-cell communication analysis by using node importance measures to compare CCC graphs.

CHAPTER 4

Learning distances from a set of cell-cell communication graphs

In Chapter 3, we revisit the initial work that used graph theory concepts introduced for CCC, based on scRNAseq data. We introduce CrossTalkeR, a tool for comparing CCC graphs using node importance measures. Moreover, we provide two translational examples in which CrossTalkeR was key to revealing communication-related biological programs in real-world applications.

With the massive volume of data generated in the personalized medicine context, patient-centric analysis has gained increasing interest in recent years (Flores et al., 2023; Joodaki et al., 2024; Armingol et al., 2022). In this Chapter, we interrogate how patient-derived CCC graphs can further reveal patterns in complex diseases (Nagai et al., 2024). Can CCI patterns be used to stratify disease substages?

To answer these complex questions, we need to learn distances across many CCC graphs. To further perform the metric learning approach, one first needs to overcome the challenges posed by the nature of CCC graph representations, particularly their non-symmetric structure. To mitigate this limitation, we evaluate and propose using node distance measures for directed weighted graphs (DWG) as an OT ground metric. Subsequently, we benchmark these measures using synthetic and real-world datasets to understand the main structural features captured by each distance metric.

In this section, we introduce the fundamental aspects of our proposed new metric for directed weighted graphs. The chapter starts with an introduction to the related literature. Next, we propose a synthetic model to simulate and demonstrate the properties captured by our OT metrics. Finally, we present the benchmarks and results for the synthetic model and real-world data.

4.1 Motivation and Related Work

As in Chapter 2, comparing graphs is a complex yet essential task. One specific way to solve such a task is by solving the graph edit distance (see Chapter 2.1), where the minimum number of edit operations for morphing a source graph to a target graph defines its distance. However, solving the exact form of GED is NP-hard and often only considers edit operations, meaning that attributes from nodes/edges are disregarded. On the other hand, distance measures that consider the algebraic representation of graphs exist. Examples include distance measures based on spectral features, the computation of specific subgraph statistics such as graphlets (Malod-Dognin and Pržulj, 2015) or graph kernels (Vishwanathan et al., 2010).

While the comparison of such graph features is powerful, it often does not elucidate precisely where one graph differs from another or how two graphs are equivalent. In recent years, OT based graph

distances (Maretic et al., 2019; Xu et al., 2019) have been developed as a possible solution for this explainability issue. In such an optimization, two graphs can then be compared by finding a transport plan (a mapping) between the two probability distributions with minimal transport cost (Peyré et al., 2019). The transport plan associated with the minimal cost can then be used as an interpretable, robust equation for the two graphs, highlighting the changes relevant to the computed distance.

Moreover, OT methods may take into account node/edge attributes when computing the required ground metric. To date, most OT-based methods for measuring network similarities have been proposed for undirected graphs (Nikolentzos et al., 2017; Chowdhury and Mémoli, 2019; Maretic et al., 2019, 2022). However, in many applications, such as in CCC, we are interested in comparing directed graphs. Yet, extending OT-based distances to directed graphs is not simple, as a *symmetric* distance metric, typically derived from the distances between nodes in the graph, is required within the cost function(s) typically employed within OT. This section will be composed of two main parts. We first introduce initial efforts to bridge OT theory and the graph comparison problem. Second, we introduce the use of node-to-node distances in directed weighted graphs as an OT cost function, a key building block for OT optimization.

4.1.1 Undirected graph optimal transport

In contrast with the graph kernel approach (Figure 4.1), optimal transport-based methods allow for measuring the similarity of graphs in a traceable manner. Nikolentzos et al. (2017) introduces a novel algorithm for comparing undirected graphs using optimal transport. To this end, each graph is represented as a collection of node embeddings (see the schematic in Figure 4.1). Graphs are then embedded in an Euclidean space using the eigenvectors of the respective adjacency matrices. Next, OT is used for computing the matching between the graph embeddings. Another approach that solves OT in undirected graphs is Graph Optimal Transport (GOT, Maretic et al. (2019)). At first, GOT considers that each undirected graph can be represented as a multivariate normal distribution in the form $\mathcal{N}(\mathbf{0}, L^{\dagger})$, where $L^{\dagger}$ is the pseudo-inverse of the Laplacian matrix. Next, GOT computes the OT mapping using the theory of multivariate Gaussian optimal transport (Takatsu, 2010). GOT uses a permutation matrix-based approach for mapping graphs with different labels.

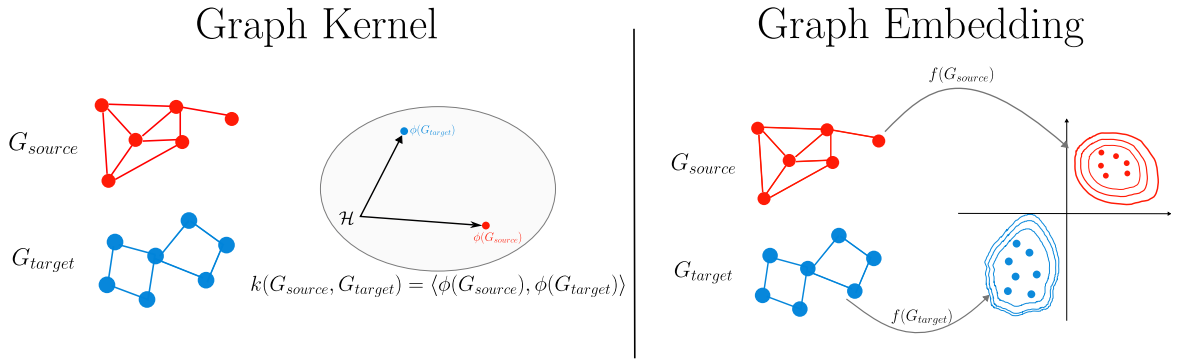


Figure 4.1: On the Left, illustration of the graph kernel rationale. On the right, an illustration of graph embeddings. Figures were inspired by Jia et al. (2022) and Togninalli et al. (2019), respectively

While the methods described above are well-defined for comparing undirected graphs, they cannot be used to compute distances on directed weighted graphs. One main reason for that is the directionality of the edges, which yields asymmetry in the graph matrices. Approaches that use a graph descriptor in an optimal transport framework exist and rely primarily on topological features of graphs, such as graphlets (Silva et al., 2023).

Initial approaches for solving OT in the case of directed graphs often are based on Gromov-Wasserstein (GW) formulations. In GWOT (Chowdhury and Mémoli, 2019; Chowdhury and Needham, 2021), each graph is considered as a metric space, and the transport problem consists of mapping each graph metric space instead of mapping the node embedding distributions. In this context, evaluating the choice of metric in OT methods is crucial for learning meaningful distances. To this end, in the chapter, we evaluate several node-to-node metrics used in the directed weighted graph optimal transport problem. The metrics introduced next are obtained considering both the direction and weights in the graphs. Laplacian operators encode the signal propagation on graphs and play a pivotal role in graph-based methods, especially for computing node-to-node distances. On the other hand, in the undirected graph case, there are several ways to define the graph Laplacian operator in the directed graph setting.

4.1.2 Graph Laplacian in Directed Weighted Graphs

Here, we introduce a set of Laplacian matrices that were developed to handle the asymmetric format of directed graphs. The introduced matrices will be of use for computing the node-to-node distances used in the directed graph optimal transport formulation proposed next. The definitions presented below will consider a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where $\mathcal{V}$ is the set of observations, $\mathcal{E}$ is the set of interactions, and $\mathcal{W}$ is the asymmetric adjacency matrix.

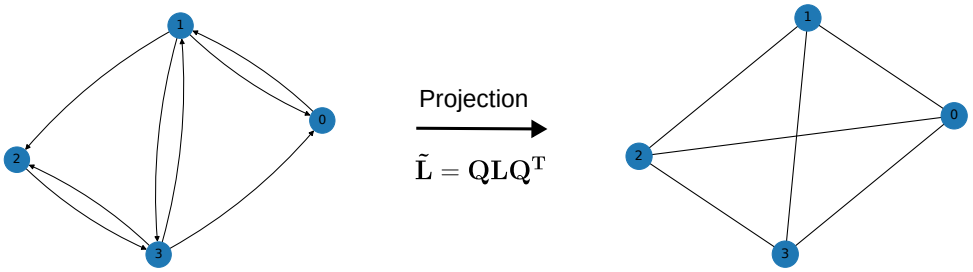


Figure 4.2: Illustration of the rationale behind how the grounded Laplacian is obtained.

Reduced Laplacian/Grounded Laplacian: Rooted in the control theory, the grounded Laplacian is proposed to solve the synchronization problem with non-identical oscillators (Jadbabaie et al., 2004; Randić and Klein, 1993). In graph theory terms, we define the Ground Laplacian (Young et al., 2015a,b) as:

Definition 4.1 (Grounded Laplacian)**.** *Let* $\mathbf{Q} \in \mathcal{R}_{N-1\times N}$ *be a (grounding) matrix which satisfies,*

$$\mathbf{Q}\mathbf{1}_N = 0 \qquad \boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}_{N-1} \qquad \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I} - \mathbf{1}\mathbf{1}^\top/N. \tag{4.1}$$

A grounded *Laplacian matrix for the graph can now be computed via Q as follows:*

$$\widetilde{\boldsymbol{L}} = \boldsymbol{Q}\boldsymbol{L}\boldsymbol{Q}^{\top}, \tag{4.2}$$

where L represents the Laplacian matrix defined in Chapter 2.2.

Magnetic Laplacian: Magnetic Laplacian is defined to keep both the Hermitian matrix properties as well as the edge directionality properties. To achieve this, the edge direction is encoded as an argument (phase) in a complex plane, as in Figure 4.3. In applications involving directed graphs, and motivated primarily by preserving Hermitian properties, Magnetic/Hermitian Laplacians have gained attention in several fields, including machine learning (Geisler et al., 2023) and computational biology (Venkat et al., 2024). Formally, the magnetic Laplacian is defined as:

Definition 4.2 (Magnetic Laplacian)**.** *A Magnetic Laplacian is a self-adjoint, positive semi-definite operator* $L^{(g)} = D - T^{(g)} \circ W^{(s)}$*, where* $W^{(s)} = \frac{\mathcal{W}+\mathcal{W}^{\mathcal{T}}}{2}$ *is the symmetric weighted adjacency matrix.* $T^{(g)}$ *is a Hermitian matrix with elements* $t_{ij} = exp(i2\pi g a_{ji})$*,* $a_{ij} = \overline{a_{ji}}$ *encodes the direction of the edges and* $0 \leq g \leq 1$ *is an electric charge parameter (Forman, 1993; Shubin, 1994; Furutani et al., 2019).*

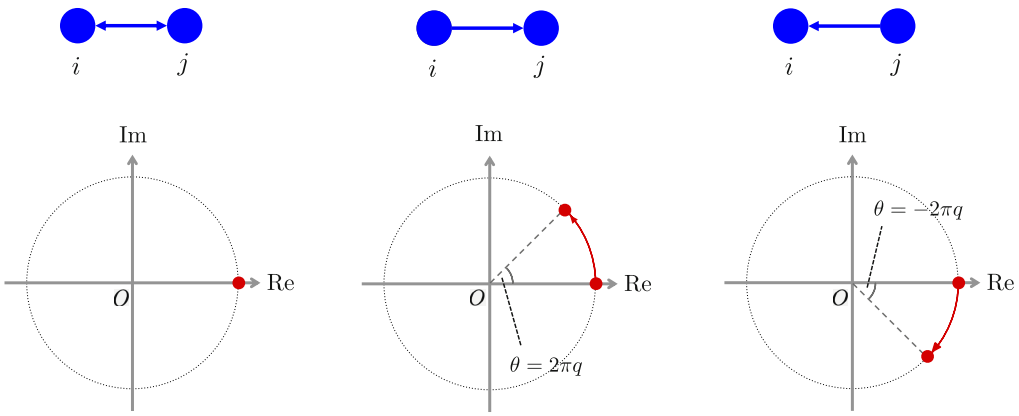


Figure 4.3: Directions enconded by the Magnetic Laplacian using $g = 1$. Adapted from Furutani et al. (2019)

4.1.3 Graph node-to-node distances

As mentioned above, a way for computing node-to-node distance using graphs is by using the Laplacian matrix. In this subsection, we introduce the resistance distance for undirected graphs. Next, we extend this computation for directed graphs. We also define other distances, which are obtained using the distinct Laplacian operators defined above.

Laplacian based node-to-node distances for undirected graphs

Definition 4.3 (Undirected node-to-node resistance distance)**.** *Let* $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ *be the adjacency matrix of a graph with N nodes, with entries* $A_{ij} = w_{ij}$ *of the edge from node i to j, and* $A_{ij} = 0$ *otherwise. The Laplacian of the graph is then given as* $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$*, where* $D = diag(\mathbf{A1})$ *is the diagonal matrix of*

(weighted) node degrees. The (square root of the) resistance distance between node i and j can then be computed as (Randić and Klein, 1993; Ghosh et al., 2008):

$$r_{ij} = \sqrt{(e_i - e_j)^\top \boldsymbol{L}^\dagger (e_i - e_j)}, \tag{4.3}$$

where $\boldsymbol{L}^\dagger$ denotes the Moore-Penrose pseudoinverse of the Laplacian, and e_i, e_j are the N dimensional indicator vectors associated to node i and node j, respectively. Note that e_i is a vector with 1 in the i-th position and 0 otherwise.

As described in Chapter 2.2, optimal transport is a powerful framework to map two distinct distributions using a given cost. Although recent methods are being developed for graphs, only a few approaches for directed graphs use node-to-node distances that account for edge directionality. While in the undirected graph case, node-to-node distances are easy to compute, in the directed graph case, several approaches attempt to compute such a matrix while preserving signal propagation and edge feature variability. A specific linear operator often used to compute node-to-node distances in undirected graphs is the graph Laplacian. An example of a distance measure that uses such a linear operator is the resistance distance (an example of this distance in a graph is available in Figure 4.4), defined in Definition4.3.

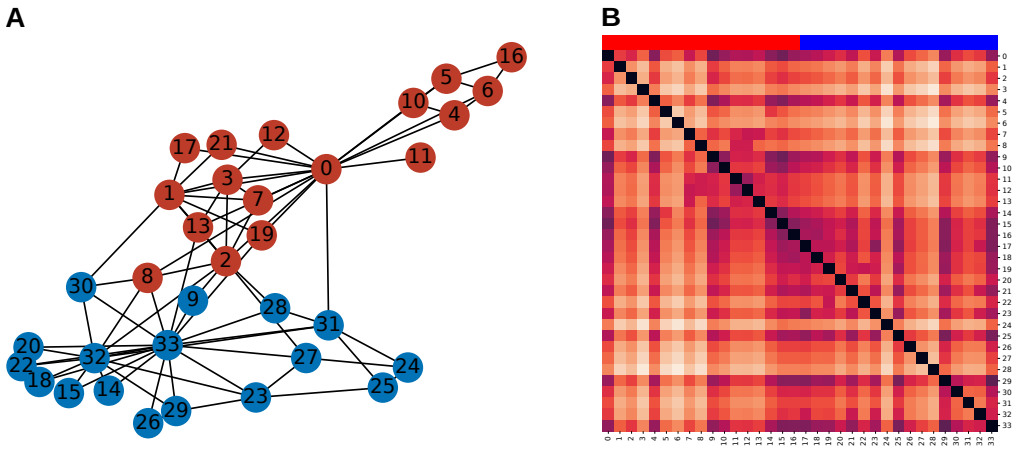


Figure 4.4: A. Karate club graph, nodes represent members, and edges are placed when members interact outside of the club activities. B Resistance Distance (as in Definition 4.3) of the club members using the graph shown in A.

Laplacian based node-to-node distances for directed graphs

In the case of directed graphs, the Laplacian matrix is a proxy for computing node-to-node distances that accounts for the edge directionality. However, there are different ways to obtain the graph Laplacian when considering directed graphs. In this section, we first introduce several methods from the literature for computing the Laplacian matrix of directed graphs. Second, we formally present the node-to-node distance that we propose as a ground metric for the directed graph optimal transport.

Generalized Effective Resistance Distance (GRD)

As in Definition 4.3, the resistance distance is a well-known distance metric for nodes within undirected networks (Randić and Klein, 1993; Ghosh et al., 2008). To generalize this notion of effective resistance to directed graphs, Young et al. (Young et al., 2015a,b) use an approach that relies on the grounded Laplacian. The GRD consists of computing the resistance distance from an undirected graph with resistances equivalent to the original directed graph, as in Figure 4.2.

Using the grounded Laplacian as defined in Definition 4.1, the (square root of) the resistance distance can now be alternatively written as $r_{ij} = [(\widetilde{e}_i - \widetilde{e}_j)^\top \widetilde{\mathbf{L}}^{-1} (\widetilde{e}_i - \widetilde{e}_j)]^{-1/2}$, where the inverse of the grounded Laplacian is used in lieu of the pseudoinverse of the standard Laplacian matrix and $\widetilde{e}_i$ and $\widetilde{e}_j$ are respectively e_i and e_j without the N-th node. Indeed, the action of the matrix Q can be understood in terms of removing the null-space of $\mathbf{L}$ by effectively "grounding" the resistor network, i.e., fixing a reference potential (Jadbabaie et al., 2004; Ghosh et al., 2008), as in Figure 4.2.

Young et al. (2015a) extended the notion of resistance distance by using an alternative characterization of the inverse $\widetilde{\mathbf{L}}^{-1}$ that appears in the definition of the effective resistance. Specifically, they define the *generalized effective resistance distance* between node k and node j by:

$$d_{\mathrm{GRD}}(k, j) = \sqrt{(e_k - e_j)^\top \mathbf{X} (e_k - e_j)}. \tag{4.4}$$

Here the matrix $\mathbf{X}$ is defined via

$$\mathbf{X} = 2Q^\top \Sigma Q, \tag{4.5}$$

where Σ is the solution of the Lyapunov equation:

$$\widetilde{L}\Sigma + \Sigma \widetilde{L}^\top = I_{N-1} \tag{4.6}$$

This solution is unique under the assumption that a globally reachable node exists within the graph. Note that this formulation reduces to the classical resistance distance for undirected graphs. For more details, we refer to Young et al. (2015a,b).

Hitting Time Based Distance (HTD) A second type of node-to-node distance metric for directed graphs is the class of hitting time metrics developed by Boyd et al. (2021).

Consider a discrete-time Markov chain $(X_t)_{t\geq 0}$ on the space of the vertices $\mathcal{V} = \{1, ..., N\}$ of a strongly connected graph, with initial distribution λ and an irreducible transition matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$ such that:

$$P(X_0 = i) = \lambda_i \qquad \text{and} \qquad P(X_{t+1} = j | X_t = i) = P_{i,j}. \tag{4.7}$$

Let $\pi \in \mathbb{R}^N$ be the invariant distribution of the chain, i.e., $\pi\mathbf{P} = \pi$. For a starting point distributed according to λ, the *hitting time* of a target vertex $i \in \mathcal{V}$ is the random variable

$$\tau_i = \inf\{t \geq 1 : X_t = i\}. \tag{4.8}$$

Following Boyd et al. (2021), we denote the probability that starting in a node i the hitting time of j is less than the time it takes to return to i by

$$Q_{i,j} := P_i[\tau_j \leq \tau_i]. \tag{4.9}$$

Based on the matrix $\mathbf{Q}$ a normalized hitting time matrix $\mathbf{T}^{(\beta)}$ can be defined in terms of its entries

$$T_{i,j}^{(\beta)} = \begin{cases} \frac{\pi_i^\beta}{\pi_j^{1-\beta}} Q_{i,j} & i \neq j, \\ 1 & \text{otherwise}, \end{cases} \tag{4.10}$$

where β is a scalar parameter that can be used to adjust the measure if $\mathbf{P}$ is an irreducible stochastic matrix (i.e., the underlying graph is strongly connected, the Hitting Time Distance Matrix for $\beta = (0.5, 1]$ can be obtained by:

$$d_{\text{HTD}}^{(\beta)}(i,j) = HTD_{i,j}^{(\beta)} \quad \text{where} \quad HTD_{i,j}^{(\beta)} = -\log(T_{i,j}^{(\beta)}) \tag{4.11}$$

Remark. (Not strongly connected graphs). Note that both distance functions discussed above make assumptions about the graph's global reachability of (all) nodes. To make the above distance measures well-defined in case these assumptions are not fulfilled, one option is to add a low-rank regularization term, as popularized within the context of the well-known PageRank algorithm (Page et al., 1998; Gleich, 2015).

Covariance Based Distance

A third type of node-to-node distance metric for directed graphs can be obtained using an empirical covariance matrix Σ. As presented in Tamaru et al. (2024), an euclidean distance matrix is defined by $Z = (I \circ \Sigma)11^T + 11^T(I \circ \Sigma) - 2\Sigma$. Analogously, the correlation distance matrix can be written as $Z = 1 - \frac{\Sigma}{\sqrt{((I \circ \Sigma)11^T)(11^T(I \circ \Sigma))}}$. In graphs, an empirical covariance is defined as the Moore-Penrose pseudo-inverse of the graph Laplacian $\Sigma = \mathbf{L}^\dagger$, with the assumption that a graph follows a Gaussian Markov Random Field model (Dong et al., 2016).

4.2 Learning metrics using directed weighted graphs

In this section, we describe our approach for learning metrics across directed weighted graphs. We first formalize how the metrics introduced before can be used in both Wasserstein-OT and Gromov-Wasserstein OT. We follow this by presenting the synthetic data we propose for benchmarking the metric learning method in a controlled manner. Next, we describe the baseline methods and the benchmark using real-world data. Finally, we present and discuss our results and future directions for the research field.

4.2.1 Directed Weighted Graph Wasserstein Distance

To derive a Wasserstein distance for directed graphs, we reinterpret the observed data, as illustrated in Figure 4.5. Instead of considering the geometry defined by each graph in our observation set $\{\mathcal{G}^1, ..., \mathcal{G}^p\}$ separately, and transporting these geometries such that the distortion is minimized, we now interpret the edge-weights of each weighted graph as a specific realization of a signal supported on the *same underlying space*. Stated differently, we consider the edge weights as a distribution supported on the same set of latent edges, and we aim to transport these edge distributions optimally

between graphs. To this end, we allocate the observed edge weights in a matrix $\mathbb{P} \in \mathbb{R}^{E \times p}$, whose rows are indexed by the (directed) edges and columns by the samples, i.e., the column $\mathbb{P}_{:,k}$ describes the weights of the kth graph $\mathcal{G}^k$.

Using this definition, we build a *directed line graph* $\mathbb{L} = (\mathbb{V}, \mathbb{E}, \mathbb{W})$ called Shared Transportation Graph. The vertex set $\mathbb{V}$ of the line graph contains each possible edge (j,k) contained in one of the graphs $\mathcal{G}^i$. The edge set $\mathbb{E}$ of the line graph is defined as follows: an edge $e' = (u', v') \in \mathbb{E}$ exists if the target of u' is the source of v', i.e., the target node of edge u' in the original graph, is the source node of edge v' in the original graph. Note that, to build the line graph, the weights of the cell-cell network interactions are not considered. This data is rather considered as a distribution supported on the nodes of the line graph. Finally, the weight $w_{u'v'}$ of edge (u', v') in the line graph is simply the proportion of graphs containing both edges u' to v'. Once we have a weighted, directed line graph $\mathbb{L}$, we can again associate a distance matrix based on the generalized effective resistance or the hitting time with $\mathbb{L}$. The Wasserstein distance between two graphs $\mathcal{G}^k$ and $\mathcal{G}^l$ can then be computed as follows:

$$d_W(\mathcal{G}^k, \mathcal{G}^l) = \underset{\Gamma \in \Pi(\mathbb{L})}{\arg\min} \quad \langle \Gamma, C_{\mathbb{L}} \rangle_F \tag{4.12}$$

$C_{\mathbb{L}}$ is one of the distance matrices (generalized effective resistance, hitting time distance) defined for directed graphs, as discussed above. Here $\Pi(\mathbb{L}) = \{\Gamma \in \mathbb{R}^{|\mathbb{V}| \times |\mathbb{V}|} : \Gamma_{ij} \leq 0, \Gamma \mathbf{1} = P_{:,k}, \Gamma^\top \mathbf{1} = P_{:,l}\}$ is again the set of all admissible transport matrices, whose marginal distribution is now, defined by the respective distribution of edge weights as encoded in the columns of the matrix $\mathbb{P}$.

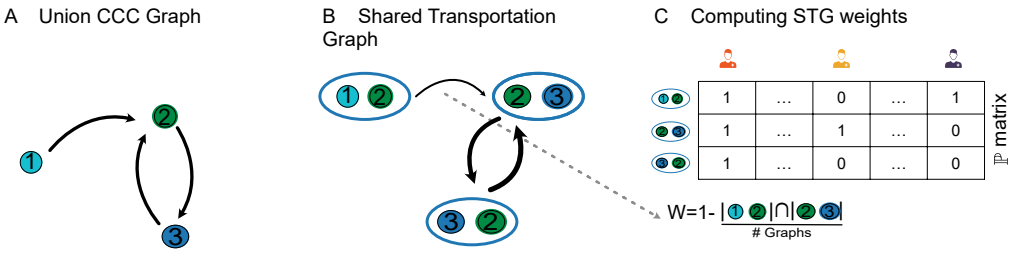


Figure 4.5: A-A union CCI graph is obtained by considering all directed interactions present in at least one sample CCI. B- The union graph is transformed to a line graph, where nodes are directed edges $1 \to 2, 2 \to 3, \ldots$. C- The weights of the edges in the line graph in B are the frequency of both interactions that appear in all sample-specific CCIs. The idea behind STD is that rarer interactions have a higher cost than more common ones.

4.2.2 Directed Weighted Graph Gromov-Wasserstein Distance

As mentioned above, to allow the computation of the Wasserstein distance, we have switched our perspective on the observations. Alternatively, as defined in Chapter 2.3, GWOT accounts for the displacement of the objects using distances in two metric spaces. This property allows us to extend the GWOT for directed weighted graphs using the node-to-node distances presented in this chapter.

To this end, we will consider our observation set $\{\mathcal{G}^1, \ldots, \mathcal{G}^p\}$. Considering the previous approaches developed in the context of undirected graphs (Chowdhury and Mémoli, 2019; Demetci et al., 2020), we can write each graph $\mathcal{G}^k$ from the set as a metric space with the form $\mathcal{X}_{\mathcal{G}^k} = (\mathcal{X}_{\mathcal{G}^k}, d_{\mathcal{G}^k}, \mu_{\mathcal{X}_{\mathcal{G}^k}})$. $\mathcal{X}_{\mathcal{G}^k}$

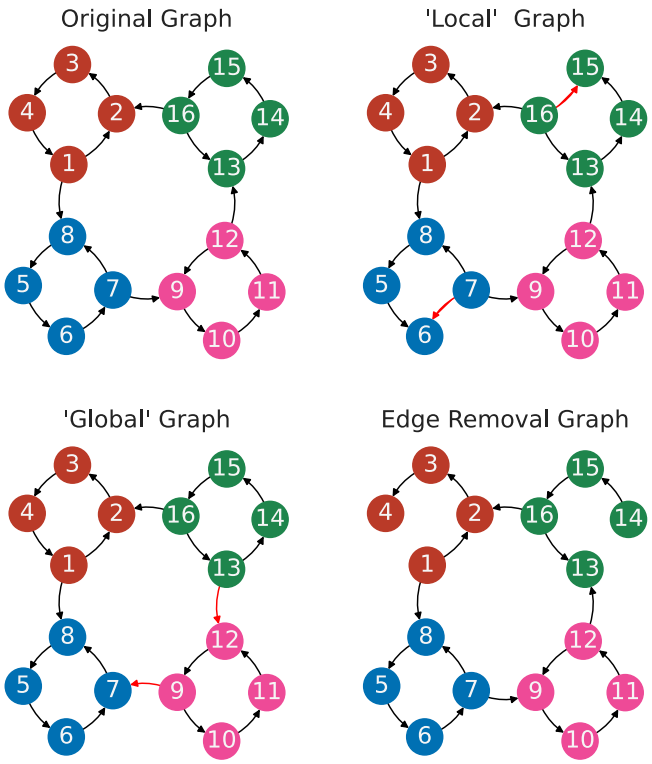


Figure 4.6: Example of graph perturbations. We manipulate the original graph by flipping an interaction (red edges) in a local cycle, named the local perturbation, or in the global cycle (global perturbation). Another perturbation is the removal of edges, which can also affect local or global cycles

can represent a feature vector in the node space, $\lceil_{\mathcal{G}^k}$ represents the node-to-node distances and $\mu_{\mathcal{G}^k}$ is the probability distribution over the nodes. To perform the GWOT approach, we considered the directed graph node-to-node distances as $\lceil_{\mathcal{G}^k}$ and considered $\mu_{\mathcal{G}^k}$ as a uniform distribution.

4.2.3 Synthetic model for benchmarking distances in graphs

To simulate the changes in topology and signal of real-world datasets, we develop a synthetic dataset based on a DWG which consists of four cyclic graphs with four nodes connected in a cyclic fashion (Rosvall et al., 2009), as in Figure 4.6. Here, our goal is to build a dataset in which changes across graphs are known and controllable.

Our dataset contains the original graph and graphs that are generated using the following perturbations:

- Flip of local Edge: edge direction of local cycle is swapped.
- Flip of global Edge: edge direction of global cycle is swapped.
- Edge removal: edge is randomly removed.

Using the perturbations above, we generated a set containing a hundred DWGs.

4.2.4 Benchmark strategies

To evaluate our approach, we used two main assays. First, the synthetic assay, where we have control of the changes across the graphs. Second, in the real-world assay, we want to evaluate how metrics can split CCC communication across conditions.

Synthetic Dataset

To start our benchmark, we examine whether our metrics are sensitive to edge weight scaling. To this end, we compared the distribution of the distances against zero using a one-sided t-test. Ideally, distance metrics should be invariant to the proposed scaling operation. Our results indicate that only measures GRD, HTD, and $dCOV_{ground}$ are scale invariant for both local and global perturbations (Figure 4.7). Interestingly, our results suggest that metrics obtained using the Gromov-Wasserstein are scaling sensitive. Large differences are presented by baseline methods.

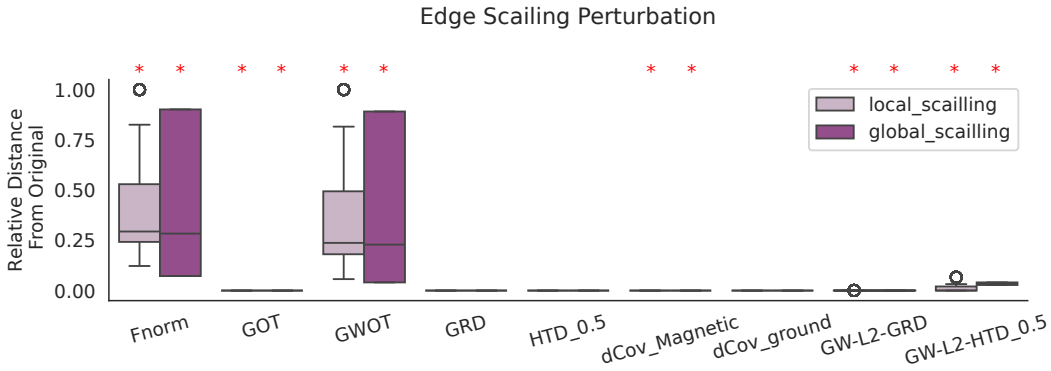


Figure 4.7: Relative distances obtained from an edge perturbation assay. (*) indicates that the distributions are different from zero (p-value < 0.0001, zero comparison t-test), i.e., they are not scale invariant.

Then, we sought to assess how methods perform when perturbing the network using the edge removal process. Here, it is expected that the graph distance increases with the number of edges removed; i.e., values for 1-edge removal should be higher than 0, and values for 2-edge removal should be higher than those for 1-edge removal. Results indicate that all distance metrics display appropriate properties. In these results, GOT, GRD, HTD, Magnetic, and Ground Laplacian display large discrimination between the distributions. Furthermore, we also compared the 1-edge removal with the zeros using a one-sided t-test. The comparison results indicate that all distribution means are higher than zero (see Figure 4.8).

Lastly, we tested our distance measures when comparing local and global perturbations in our synthetic model(Figure 4.9). We compare the distance distributions using a one-sided Wilcoxon test to determine whether the median of the local perturbation is higher than that of the global perturbation. Surprisingly, our results indicate that the baselines (Fnorm, GOT, GWOT) are unable to identify such differences. GRD, HTD, the $dCov_{Magnetic}$, and the GW distances identify that local perturbations

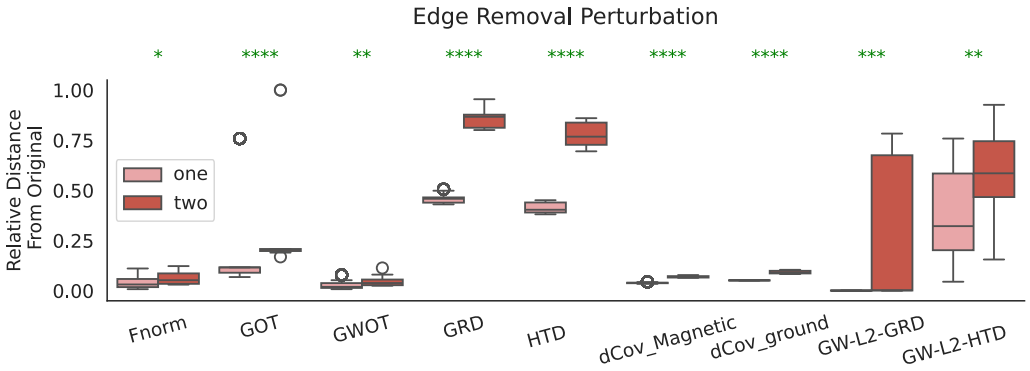


Figure 4.8: Relative distances obtained from an edge perturbation assay. The comparison between the perturbation distributions. For one edge removal, the * indicates distributions significantly distinct than zero (Wilcoxon test, p-value < 0.05). For two-edge removal, * indicates distributions higher than single-edge removal (Wilcoxon test; p-value < 0.05).

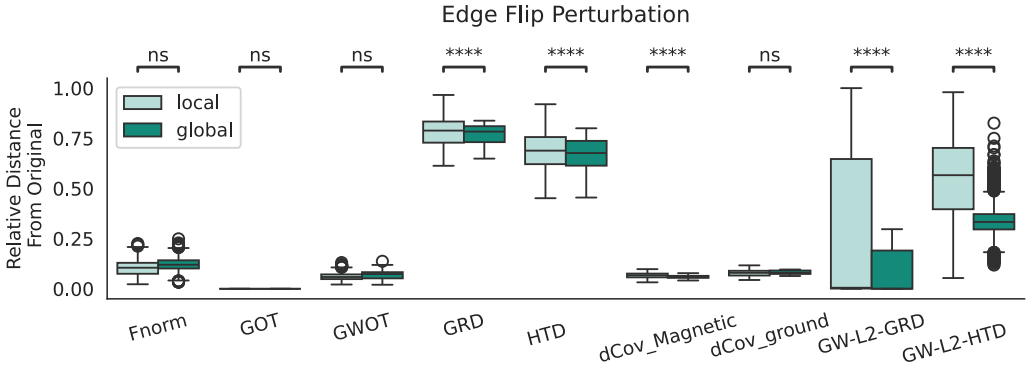


Figure 4.9: Relative distances obtained from an edge perturbation assay. The comparison between the perturbation distributions. For one edge flip, the * indicates that the means from the distribution are different from zero (Wilcoxon Test, p-value < 0.05).

induce higher distances than the global perturbations. We hypothesize that this is mainly because local perturbations can isolate individual nodes.

To further investigate the reasons for such a difference, we computed the number of zero indegree(source)/outdegree(sink) induced per perturbation. As indicated in Figure 4.10, we observe that global perturbations do not result in sources or sinks in the graphs. Overall, it is possible to hypothesize that local perturbations increase the number of sinks and sources, which may explain differences in the distance distributions. The presence of sink and source also results in abrupt flow changes once nodes can be isolated from the general flow.

In summary, our results indicate that Wasserstein OT distances overperform the Gromov-Wasserstein distances. Regarding cost functions, the best results are achieved when using the GRD, $HTD_{0.5}$, and $dCoV_{ground}$ cost functions. Additionally, our metrics were able to identify several kinds of perturbations. Perturbations where sink/source are formed tend to present higher distances than perturbations where the information flow is unaltered.

Remark. (Relative Distance). Note that we have considered the relative distance metric to compare

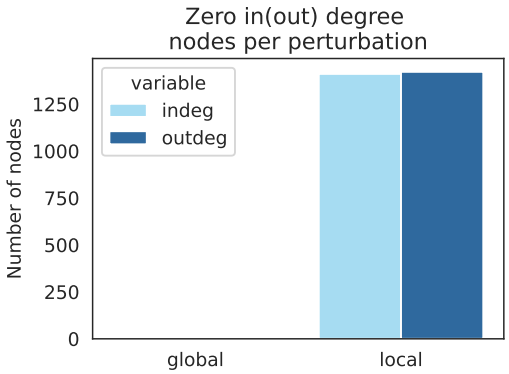


Figure 4.10: Number of sinks(zerod-outdegree) and sources(zerod-indegree) found in the perturbed graphs.

the distances. This is obtained by normalizing the OT distance by its maximum distance.

Real-World Application

To illustrate the utility of our graph distances, we further consider a real-world analysis task, namely the comparison of cell-cell communication networks. Specifically, we consider a scenario in which we are given a patient cohort with p patients, from which we generate a set of cell-cell communication networks $\mathbf{G} = \{\mathcal{G}^1, ..., \mathcal{G}^p\}$ using the ligand-receptor analysis method CrossTalkeR (Nagai et al., 2021), as described in Chapter 3.

Our main question is whether comparisons of cell-cell networks can be used to delineate disease stage or relevant subgroups in these patients. To construct our evaluation dataset, we retrieved ten publicly available single-cell RNA sequencing datasets. For each dataset, we generated the patient graphs as described before. Each graph is labeled according to the patient status in the cohort. Next, we compute graph-to-graph distances using the distance measures we introduced.

To assess how far we can find a partition of the patient cohorts, we perform two evaluation approaches: (1) clustering-based (Unsupervised) and (2) classification-based. When benchmarking using the clustering-based approach, we perform k-medoids with the estimated distance matrix as input. For simplicity, we set the number of clusters to be exactly as the number of classes in the data and use the Adjusted Rand index (ARI) (Hubert and Arabie, 1985) between the true label and clustering result to evaluate the performance of the clustering result. Orthogonally, when benchmarking using the classification approach, we also assess how patients are classified using the distinct distance matrices. This evaluation was performed in an 80/20 fold cross-validation with a k-nearest neighbors classifier, considering $k = 3$ nearest neighbors.

We also include baseline methods. The baseline is a simple approach for measuring distances from the matrix $\mathbb{P}$. Here, the matrix $\mathbb{P}$ is used as an input for a Principal Component Analysis, which can be obtained by computing an SVD of the centered $\mathbb{P}$ matrix $\tilde{\mathbb{P}} = U\Sigma V^\top$. Next, the principal components matrix $U_{p\times(p-1)}$ is used to compute the pairwise Euclidean distance between the samples. For the clustering problem, HTD and $dCov_{ground}$ obtain the highest average values. For the KNN classifier, the $dCov_{ground}$ obtains the highest accuracy. As expected, the baseline approaches (PCA), which

Table 4.1: K-Medoids Adjusted Rand Index

	PCA	GRD	$HTD_{0.5}$	$dCov_{Magnetic}$	$dCov_{ground}$	GW_{GRD}	GW_{HTD}
Breast Cancer	-0.0122	0.2770	**0.2770**	**0.2770**	**0.2770**	0.1358	0.1194
scRNAseq Kidney AKI/CKD	-0.0275	0.0797	**0.1402**	0.0797	**0.1402**	-0.0371	0.1290
Myocardial Infarction	-0.0087	**0.3129**	**0.3129**	**0.3129**	**0.3129**	-0.0123	0.0850
Lung Adenocarcinoma	-0.0113	0.1478	**0.1633**	0.1478	0.1627	0.0580	0.0594
COVID	0.0484	0.2236	0.2236	0.2236	0.2236	**0.4156**	0.3361
snRNAseq Kidney AKI/CKD	-0.0221	-0.0259	-0.0259	-0.0259	-0.0259	-0.0425	**0.1192**
COVID2	0.0062	0.0053	**0.0102**	0.0085	0.0053	0.0008	-0.0000
Diabetes	-0.0421	0.0950	0.0950	0.0950	0.0950	0.0313	**0.1686**
Lupus	-0.0005	**0.0586**	**0.0586**	**0.0586**	**0.0586**	-0.0064	0.0154
Pancreatic Cancer	0.2160	**1.0000**	**1.0000**	**1.0000**	**1.0000**	-0.0075	0.0149
Mean	0.015	0.217	**0.225**	0.218	**0.225**	0.054	0.105
Std. Dev.	**0.054**	0.297	0.294	0.297	0.294	0.138	**0.099**

do not account for direction, were among the worst performers in both tasks. Moreover, the GW underperformed in the clustering task.

Altogether, these results indicate that the Wasserstein, coupled with the node-to-node covariance, is better suited for sparse and noisy DWG obtained from cell-cell communication networks from single-cell RNA sequencing. A potential explanation may be that the Wasserstein formulation is adaptive with respect to the whole ensemble of graphs considered due to the specific construction of the line graph. Furthermore, by considering the covariance distance, we obtain a smoother distance, improving both clustering and classification tasks.

Table 4.2: KNN Classifier Mean Accuracy

	PCA	GRD	$HTD_{0.5}$	$dCov_{Magnetic}$	$dCov_{ground}$	GW_{GRD}	GW_{HTD}
Breast Cancer	0.6923	0.7500	0.7760	0.7913	**0.7927**	0.7832	0.7721
scRNAseq Kidney AKI/CKD	0.4750	0.6188	0.6719	0.7141	**0.7194**	0.6989	0.6813
Myocardial Infarction	0.6000	0.8000	0.8438	0.9219	0.9306	0.9341	**0.9187**
Lung Adenocarcinoma	0.4182	0.5318	0.5750	0.6076	**0.6121**	0.5928	0.5760
COVID	0.5962	0.7462	**0.8260**	0.8620	0.8658	0.8538	0.8397
snRNAseq Kidney AKI/CKD	0.3750	0.4250	0.4062	0.4406	0.4444	0.4432	**0.4604**
COVID2	0.3387	0.3710	0.3919	0.4008	0.4025	**0.4041**	0.3946
Diabetes	0.2909	0.5091	0.5955	0.6648	**0.6717**	0.6669	0.6636
Lupus	0.5415	0.6783	0.7401	0.7781	**0.7818**	0.7660	0.7549
Pancreatic Cancer	0.5429	0.7714	0.8857	**0.9429**	0.9492	0.9286	0.9000
Mean	0.487	0.620	0.671	0.712	**0.717**	0.707	0.696
Std. Dev.	**0.129**	0.153	0.176	0.187	0.188	0.185	0.177

4.3 Discussion

Comparing graphs can help explain key changes in interactions across different system stages. For undirected graphs, several approaches have been introduced in recent years. Developed methods can prioritize not only the topological identity of graphs but also similarity based on shared node/edge features. On the other hand, for directed graphs, only a few methods can account for the directionality of interactions. This fact is mainly due to the asymmetric nature of directed graphs.

Among the solutions available for comparing graphs, graph optimal transport stands out for providing an interpretable mapping via the Transport map. The transport map can be used to trace how the compared graphs are mapped. As we have shown in Chapter 2, in OT we seek a mapping between two input distributions, given a cost (i.e., a ground metric). Within this context, we developed a graph optimal transport method for directed weighted graphs, and our main contributions are:

- We introduce a synthetic dataset containing several perturbations
- We propose and benchmark node-to-node distances that account for the edge directionality as OT ground metric options
- We benchmark the graph-to-graph distances for comparing CCC graphs.

In the realm of comparing directed graphs, there are only a few synthetic datasets in which both directionality and edge weights are considered. Moreover, often benchmark graph datasets are complex, and the changes across graphs are not traceable. To overcome this, we elaborate a directed weighted dataset containing three main perturbation types, namely, edge directionality swap, edge weight scaling, and edge removal. Our dataset was used to compare state-of-the-art methods for comparing graphs and the methods developed in this chapter. By examining the distribution changes in the learned graph distances, it is possible to identify the distances that are able to better discriminate the DWG's changes. Moreover, we can show that distances are overall sensible for identifying sources (zero in-degree)/targets (zero out-degree) nodes in the graphs.

Our OT methods use three distinct node-to-node distances as ground metric options. All options were obtained by using the directionality encoded by Laplacian matrices. Our benchmark shows that Wasserstein-OT outperforms Gromov-Wasserstein methods, suggesting that GW may be sensitive to noise when computing directed node-to-node distances across graphs. Moreover, the choice of ground metric also affects the discriminative power of the Optimal Transport metric. Within the evaluated metrics, for both synthetic and real-world data, benchmarked HTD and $dCOV_{ground}$ presented the best performance. We also notice that HTD is able to better recover perturbations on directionality than $dCOV_{ground}$.

Here, we proposed OT methods for comparing directed weighted graphs. Results from both synthetic and real-world data indicate that Wasserstein-OT is the most effective for comparing directed weighted graphs. Considering the edge directionality in the OT cost function is beneficial for comparing graphs in the context of cell-cell communication.

Our results also pave the way for further improvements, both computationally and biologically. On the computational side, scalability with respect to dataset size and the number of nodes still needs to be investigated. Moreover, the use of graph kernels that consider edge attributes can also be an alternative approach for comparing graphs. On the biological side, leveraging the OT-metric space for downstream tasks, such as clustering, can help reveal application-related patterns of interest.

CHAPTER 5

Clustering of cell-cell communication graphs

As introduced in the previous chapters, this thesis encompasses the graph comparison problem in the context of cell-cell communication. In Chapter 2, we provide the basic knowledge for understanding the thesis. Chapter 3 focuses on the usage of graph node importance measures for comparing graphs in a pairwise manner. We developed CrossTalkeR for prioritizing and comparing CCC graphs of different diseases, as an example of CCCs of diseased and non-diseased samples.

Building on our graph comparison endeavor and motivated by the rapid development of sequencing technology, in Chapter 4, we further investigate how to compare CCC graphs from patients. To answer this question, we use node-to-node distances that account for CCI's orientation and weights as a ground metric for OT-based methods. Using the OT metric, initial results indicate that it can discriminate patients by disease stage.

Here, we further investigate the use of graph-to-graph distances to identify subgroups of diseased samples (Nagai et al., 2025). To this end, we developed a computational approach that clusters graphs using the learned graph-to-graph distances. This chapter begins with a background and related work section. We next introduce our novel approach to directed weighted graph clustering. This is followed by benchmarks for clustering, robustness, and patient survival. Next, we introduce our case study, which uses a pancreatic cancer dataset. At last, we discuss the remaining main challenges, our contributions to the field, and future directions for the CCC graph clustering problem.

5.1 Motivation and Related Work

There are now hundreds of computational methods for LR-based communication analysis (Armingol et al., 2024). These tools mainly focus on inferring LR pairs within a *single* biological condition. As pointed out in Chapter 2, a yet poorly studied aspect is the characterization of changes in cell communication across *multiple* biological conditions, such as disease vs. control or during cell differentiation (Cesaro et al., 2025). Here, we recapitulate two existing approaches for handling patient data in the CCC context.

5.1.1 Tensor-Cell2Cell

Tensor-cell2cell (Armingol et al., 2022) used a tensor factorization approach to reveal CCC patterns in scRNAseq data. Tensor-cell2cell takes as input a list of ligand-receptor pairs and an annotated scRNAseq gene expression matrix. This input is then used to build a tensor representation of the data,

which is related to each context via a tensor decomposition. Here, each context can be considered a patient or a biological condition.

Context-communication tensor assembly

The communication tensor assembly has three main steps. First, a communication matrix is built for each LR pair from the interaction list. To this end, a communication score is assigned to each LR pair for each source-target cell pair. Examples of such a communication score are described in Chapter 2. Second, the LR pair matrices are aggregated in a 3-dimensional communication tensor with the dimensions being LR pair × source × target. The third and final steps consist of generating the tensors as in Steps 1 and 2 for all contexts (i.e., samples). The final result is a 4-dimensional tensor (sample × LR pair × source × target).

Tensor decomposition analysis.

After obtaining the communication tensor, Tensor-cell2cell performs a non-negative tensor component analysis (TCA). TCA is a generalization of the non-negative matrix factorization (NMF) for high-order tensors. In mathematical terms, let $X_{C\times P\times S\times T}$ be a tensor with C,P,S, and T being the number of samples, number of LR pairs, number of sources, and targets, respectively. X_{ijkl} represents the interactions of samples i via the LR pair j between the source cell k to the target cell l. The TCA method is done via the CANDECOMP/PARAFAC implementation and yields the decomposition, factorization/approximation of X using a sum of R rank-1 tensors:

$$X \sim \sum_{r=1}^{R} c^r \otimes p^r \otimes s^r \otimes t^r, \tag{5.1}$$

where $\otimes$ represents the outer product operation and c^r,p^r,s^r, and t^r are vectors of the factor r that contain the loadings of the respective elements in each tensor dimension. Note that vectors have values greater than or equal to zero. In a similar fashion to the NMF, elements with large loading values have a higher importance in the biological pattern captured by the corresponding factors, and factors are permutable.

The tensor factorization is performed iteratively, following the objective function until convergence. This iterative process is performed using an alternating least squares minimization.

$$\min_{\{c,p,s,t\}} \left\| \mathcal{X} - \sum_{r=1}^{R} c^r \otimes p^r \otimes s^r \otimes t^r \right\|_F^2, \tag{5.2}$$

where $\|\bullet\|_F^2$ is the squared Frobenius norm of a tensor. This can be computed as:

$$\|\mathcal{X}\|_F^2 = \sum_{i=1}^{C}\sum_{j=1}^{P}\sum_{k=1}^{S}\sum_{l=1}^{T} {\mathcal{X}_{ijkl}}^2 \tag{5.3}$$

.

Finally, analogous to estimating the unexplained variance in a principal component analysis, Armingol et al. (2022) provides a function for estimating the tensor reconstruction error. This is computed

using the following equation:

$$\frac{\left\|\mathcal{X} - \sum_{r=1}^{R} c^r \otimes p^r \otimes s^r \otimes t^r\right\|_F^2}{\|\mathcal{X}\|_F^2}. \tag{5.4}$$

By obtaining contributions from each component via the TCA, Tensor-cell2cell identifies latent factors of cell–cell communication and their changes across samples (e.g., across subjects with different disease severity, multiple time points, and different tissues). Each factor captures the contribution of each communication variable and serves as a proxy for cell types and sample variables (e.g., disease stages).

5.1.2 MultiNicheNet

MultiNicheNet (Browaeys et al., 2023) is a computational tool for studying differences in intercellular and intracellular communication across multiple conditions from multi-sample scRNAseq data. MultiNicheNet leverages the results of NicheNet for sample-level analysis (Browaeys et al., 2020). To achieve this, MultiNicheNet combines scores from four distinct and complex analysis, as describe in Chapter A.1, namely, (1) the differential expression of the ligand and its cognate receptor(s); (2) the ligand activity in the receiver celltype; (3) cell-type and condition specific expression of the ligand and its cognate receptor(s) and; (4) the proportion of the LR expression in samples of given conditions. MultinicheNet can identify differential and cell-type-specific communication between cells across various samples with complex experimental designs. Furthermore, MultinicheNet employs robust statistical methods to adjust for experimental batch effects and covariates. Altogether, MultiNicheNet reveals multi-sample and multi-condition cell-type-specific differentially expressed ligand receptor pairs.

5.1.3 Limitations of Current Tools

To this date, only two computational methods for CCC – Tensor2Cell and MultiNicheNet – have considered data from multiple samples (patients). MultiNicheNet (Browaeys et al., 2023) builds upon NicheNet (Browaeys et al., 2020), considering both extracellular and intracellular signaling in CCC. To evaluate various samples, MultiNicheNet obtains pseudo-bulk representations by grouping cells by cell type and sample, and then performs a multiconditional differential communication analysis using a differential expression approach. However, MultiNicheNet is a supervised algorithm that requires the group of samples to be defined in advance and thus does not allow the identification of unknown groups of patients with distinct CCC programs. TensorCell2Cell (Armingol et al., 2022) uses tensor component analysis to detect latent factors explaining changes in CCC associated with sample-level scRNAseq data. The factors can detect patterns (CCC events) related to individual samples. Similar to MultiNicheNet, Tensor2Cell does not provide any approach for finding unknown groups of samples defined by distinct CCCs in an unsupervised manner.

5.2 Clustering Methods via Expectation Maximization

Here, we revisit and formalize the main expectation-maximization (EM) clustering approaches. These concepts will serve as the basis for the algorithm we developed, and they are described next.

Before we describe the EM algorithm, let S be the set of observed data, Z a set of latent variables (i.e., missing values) and Θ a vector of unknown parameters. The maximum likelihood estimate (MLE) of Θ is obtained by maximizing the following:

$$L(\Theta; X) = \int_p (X, Z|\Theta) dZ \tag{5.5}$$

$$L(\Theta; X) = \int_p (X|Z, \Theta) p(Z|\Theta) dZ, \tag{5.6}$$

$L(\Theta; X)$ is also known as marginal likelihood. It is important to notice that Z is unknown, meaning that computing the exact solution of $L(\Theta; X)$ is intractable. Given this intractable problem, EM algorithms aim to find the maximum-likelihood estimate in the presence of missing/unobserved data. In short, EM aims to find $\hat{\Theta}_{MLE}$ such that $\hat{\Theta}_{MLE} = \arg\max_{\Theta} L(\Theta; X)$. To solve this optimization problem, EM iterates in two steps: the expectation step (E-step) and the maximization step (M-step).

In the E-step, an expected value of the log-likelihood function of Θ, $Q(\Theta|\Theta^t)$, is computed. This estimate is obtained with respect to the current conditional distribution of Z given X and the current estimates of Θ^t:

$$Q(\Theta|\Theta^t) = E_{Z\ p(\cdot|X,\Theta^t)}\left[log p(X, Z|\Theta)\right]. \tag{5.7}$$

The M-step is computed after the E-step and finds the parameters to maximize

$$\hat{\Theta}^{t+1} = \arg\max_{\Theta} Q(\Theta|\Theta^t). \tag{5.8}$$

A classic example of the EM method for clustering is the Gaussian Mixture Model (Dempster et al., 1977).

5.2.1 Wasserstein Barycenters

Moreover, as presented in Chapter 4, optimal transport is a computational framework that possesses interesting properties, such as the transportation maps and the barycenters. Given a set of probability distributions, an OT barycenter is a probability distribution equidistant from all other distributions with respect to the Wasserstein distance. Written differently,

Definition 5.1 (Wasserstein Barycenters[Agueh and Carlier (2011)])**.** *A Wasserstein barycenter of N probability measures* $(\nu_1, \cdots, \nu_N)$ *in* $\mathbb{P} \in P(\Omega)$ *is a minimizer of* f *over* $\mathbb{P}$ *in which,*

$$f(\mu) = \frac{1}{N} \sum_{i=1}^{N} W_p^p(\mu, \nu_i) \tag{5.9}$$

The OT barycenter can have many interesting applications, for example, as a summary measure for an EM clustering algorithm Cuturi and Doucet (2014).

5.3 scACCorDiON: Clustering Patient Cell-Cell Communication Graphs

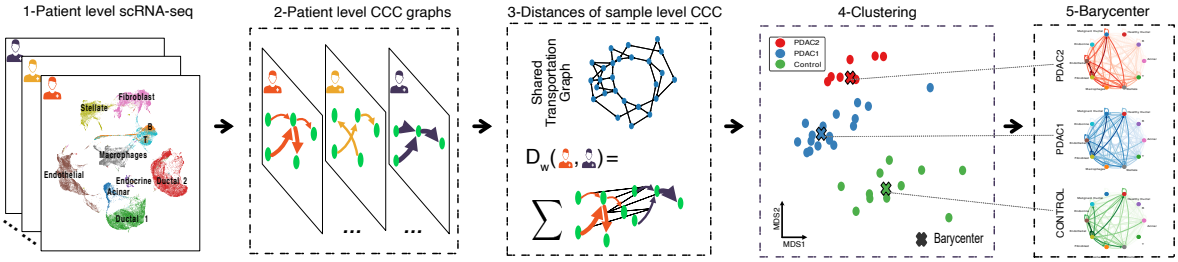


Figure 5.1: scACCorDiON receives a scRNAseq cohort experiment as input. In 1 and 2, ligand-receptor analysis is performed, and the cell-cell communication graphs for every sample (patient) are recovered. Next, in 3, we use graph-based optimal transport, which considers directions and weights of cell-cell interactions, to measure the distance between all pairs of samples. Lastly, 4 and 5, this distance is used as input a–for k-medoids or k-barycenter algorithms that find groups of CCC graphs and produce representative CCC graphs for different patient conditions, and b–for dimension reduction algorithms to create low-dimensional data visualizations. In summary, scACCorDiON enables the analysis of patient cohorts at a CCC graph level and allows for a quantitative comparison of the changes in CCC graphs using optimal transport. Moreover, Barycenters/Medoids are a proxy to facilitate the explainability of the produced results.

As described above, currently, there are no tools able to analyse CCC patient data in an unsupervised manner, and more specifically, no tools for clustering patients using CCC are available. A byproduct of the OT optimization is the barycenter, a probability distribution equidistant from all others with respect to the Wasserstein distance.

To further improve the OT framework for distances for CCC graphs (as in Chapter 4) with a barycenter-based EM approach, we propose scACCorDiON (single-cell Analysis of Cell-Cell Communication in Disease clusters using OT in directed Networks). scACCorDiON is an OT-based framework for directed weighted graph metric learning and clustering. For this, we assume that the probability masses transported via OT correspond to the directed cell-pair interaction strength (LR expression values). scACCorDiON adopts a balanced OT approach, i.e., it considers that the "mass" of cell-cell communication signals is conserved between patients and that the same amount of cell-cell communication is present in both standard and disease samples. To model this, we lift each CCC graph to a line graph, where a node represents a directed interaction, and its mass (or weight) represents the LR expression values of this interaction.

As described in Chapter 4, scACCorDiON encodes interaction information in two ways. First, it uses a cost function for OT, the Hitting Time Distance (HTD) (Boyd et al., 2021), which is obtained by

treating the directed graph as a Markov chain. Moreover, in a line graph, the masses to be transported are related to a directed cell-pair interaction.

We use two clustering approaches based on the estimated distance matrices: a k-means algorithm (Lloyd, 1982) and a barycenter algorithm. For the latter, we take advantage of the fact that OT enables us to estimate the barycenters of a set of CCC networks (Cuturi and Doucet, 2014). These barycenters can be used as "centroid" values within an expectation-maximization clustering algorithm, denoted k-barycenters. Also, barycenters can be used in conjunction with transport maps for interpretation, i.e., to delineate cell communication events that change across groups of samples.

The input to scACCorDiON is a set of directed weighted CCC graph $\{\mathcal{G}^1, ..., \mathcal{G}^p\}$ containing p graphs. We assume that each of these p graphs has been obtained from a ligand-receptor analysis method as described in Chapter 3, applied to a scRNAseq dataset containing multiple samples (e.g., a cohort). The CCC graph of sample k is then defined as $\mathcal{G}^k = (V^k, E^k, w^k)$, where a node $v \in V^k$ represents a cell type, and a directed edge $e \in E^k$ connects a pair of cell types when these cells are predicted to be communicating through a ligand (source cell) and receptor (target cell) pair. The weights of the edges w^k are related to the amount of communication between the source and target cell, e.g., $w^k(e)$ is the sum of all ligand-receptor expressions (LRScore). Note that in our problem setup, nodes can be identified across all graphs for a given scRNAseq dataset, i.e., all samples k have the same cell types and thus the same node-set $V^k = V$.

5.3.1 Clustering Patients' Networks

scACCorDiON employs the metric d_W to perform clustering of CCC graphs. One approach for this is a K-medoids partitioning algorithm (Kaufmann and Rousseeuw, 1987; Schubert and Rousseeuw, 2019, 2021), which only requires a distance matrix and detects samples (medoids) as representative of clusters. scACCorDiON leverages the fact that we can also compute barycenters for distributions of (directed, weighted) graphs via the Wasserstein optimal transport framework (Cuturi and Doucet, 2014).

K-barycenters clustering

A Wasserstein barycenter of a set of graphs $\mathbf{G} = \{\mathcal{G}^1, ..., \mathcal{G}^p\}$ can be defined as:

$$barycenter(\mathbf{G}) = \arg\min_{\mu} \frac{1}{p} \sum_{i=1}^{p} d_W(\mathcal{G}^i, \mu(i)). \tag{5.10}$$

Time-efficient solutions to this problem can be obtained by using a dissimilarity-based loss function (Sinkhorn) of the optimal transport algorithm (Cuturi and Doucet, 2014). In addition to that, we use barycenters to define an expectation-maximization-based clustering algorithm, where barycenters represent the "centroids" and we use the Wasserstein distance (Eq. 4.12) between graphs and barycenters.

Given k as the number of desired clusters, Y be an indicator variable, where $y_i \in \{1, ...k\}$ indicates the cluster of $\mathcal{G}^i$, and $\{\mu^1, ..., \mu^k\}$ indicates the set of barycenters, this leads to clustering algorithm (algorithm 5.1). To avoid local maxima due to the random initialization, we repeat the optimization process 100 times and select the solution with the lowest average Wasserstein loss per cluster. Moreover, we

Algorithm 5.1: K-Barycenters EM

Input: CCC graphs $\{\mathcal{G}^1,\ldots,\mathcal{G}^p\}$ and number of clusters $k \in \mathbb{N}$
Output: $y = (y_1,\ldots,y_p)$ where $y_i \leq k$
// Initialize barycenters
$(\mu_1,\ldots,\mu_k)$ **repeat**
  // E-step: responsibilities
  $y_i = \arg\min_{j=1,\ldots,k} d_w(\mathcal{G}^i,\mu_j), \qquad \forall i = 1,\ldots,p$
  // M-step: update parameters
  $\mu_j = \text{barycenter}\Big(\{\mathcal{G}^i \mid y_i = j\}\Big), \qquad \forall j = 1,\ldots,k$
until *convergence of log-likelihood*;
return $\{\pi_j,\mu_j,\Sigma_j\}_{j=1}^k$;

use a seeding process to pick the initial barycenters based on selecting observations that maximize the cluster-to-cluster distance as described in the k-means++ manuscript (Arthur et al., 2007).

Data availability

scACCorDiON code is available at https://github.com/CostaLab/scACCorDiON/ and https://scaccordion.readthedocs.io/en/latest/ and the survival analysis can be found at https://github.com/CostaLab/scACCorDiON.su. The data and results are available upon request at https://zenodo.org/doi/10.5281/zenodo.10808382

5.4 Benchmark

For benchmarking, we have collected seven publicly available disease scRNAseq cohorts, from which samples were annotated with their disease status. We obtained preprocessed, integrated, clustered, and annotated objects for all datasets from (CZI Single-Cell Biology et al., 2023). An exception is the pancreas adenocarcinoma dataset, which I preprocessed and described in Joodaki et al. (2024). For the details of the considered datasets, see Table 5.1

Table 5.1: Benchmark Dataset Description
Main features of data sets used in the benchmark, including the number of cell types (Cells), the average number of directed cell-cell interactions (CCI) detected with ligand-receptor analysis, number of individuals/samples, number of cells, and number of sample labels.

Data / Study	Cells	CCIs	Samp.	No. Cells	Label	References
Pancreas Adenocarcinoma	10	75	35	57.530	2	(Peng et al., 2019)
COVID	18	245	130	647.366	4	(Stephenson et al., 2021)
Myocardial Infarction	33	871	23	132.888	3	(Kuppe et al., 2022)
Breast Cancer	10	94	126	714.331	2	(Kumar et al., 2023)
Kidney AKI	13	142	36	76.020	3	(Lake et al., 2023)
Lung Atlas	12	117	165	941.504	5	(Sikkema et al., 2023)
RCC	40	1280	17	50.236	2	(Zvirblyte et al., 2024)

The LR inference was performed with CellPhoneDB (Efremova et al., 2020) implemented in the LIANA (Dimitrov et al., 2022) framework by only considering cells in a patient sample. The parameter related to the minimum expression proportion for the ligands/receptors is set to $exp_prop = 0.15$, and highly significant interactions were considered $p - value \leq 0.01$. A description of the data set's main features is provided in Table 5.1. CCC graphs were obtained as described in Chapter 3. While scACCorDiON can be used with any LR inference algorithm, we chose CellPhoneDB due to its widespread use in the literature. We also note that CellPhoneDB is the best-performing tool in the recent single-cell benchmark (Luecken et al., 2024).

To access several characteristics of scACCorDiON results, we benchmarked scACCorDiON in several ways. First, we proposed and compared scACCorDiON with baseline approaches. Second, as in Chapter 2, scRNAseq can be viewed as a snapshot of cellular gene expression and can reveal cellular identity at multiple levels of resolution. To assess scACCorDiON's robustness to the cell-type identification process, we performed a robustness analysis using a dataset for both sampling and cell-type identification bias. At last, results obtained from scACCorDiON's case study were validated using an orthogonal dataset. Here, we used the Cancer Genome Atlas (TCGA) to perform a novel LR-based survival analysis.

5.4.1 Benchmarking cell-cell communication graph clustering

We are unaware of another computational approach that can cluster samples by considering CCC information. However, we can contrast scACCorDiON with the following baseline approaches. First, we consider tabular representations (edge weight matrix $\mathbb{P}$) of the data as input (`Tabular`). Due to the high dimensionality of $\mathbb{P}$, we first perform a dimension reduction with Principal Component Analysis (PCA). We compute the correlation distance on the matrix $\mathbb{P}$ as an OT baseline method and cost function for the previously described OT framework. This approach is denoted CORR-OT. Note that the last approach does not consider the graph's directions. We also included an undirected graph optimal transport method in our benchmark, `GOT` (Maretic et al., 2019). GOT receives a single graph as input for every patient, where two cell types are connected with the cell pairs detected in one of the directions. The average LR scores (one for each direction) provide the edge weights.

For distance metrics obtained by evaluated methods (`Tabular`, `GOT`, `CORR-OT` and `DW-OT`), we run k-barycenters, k-medoids and the community detection algorithm Leiden (Traag et al., 2019). The last algorithm is chosen based on its widespread use in scRNAseq pipelines (Wolf et al., 2018). Note also that for `Tabular`, we use k-means algorithm as this is equivalent to a k-barycenter in an Euclidean space. Algorithms were run by varying the number of clusters from 2 to 7. The Adjusted Rand Index (Hubert and Arabie, 1985) (ARI) and Rand Index (Rand) for the k equal to the number of class labels and k with maximum ARI value were computed for each clustering. Here, the disease labels are used as true classes. The Friedman-Nemenyi post-hoc test was used for every metric, clustering, and distance combination to statistically address the rank differences (Nemenyi, 1963; Demšar, 2006). For Leiden (Traag et al., 2019), we vary the resolution parameter from 0 and 1 with 0.01 steps, as this allows us to obtain distinct clusters. We refer to Supp. Fig. A.1 for an overview of the experimental design.

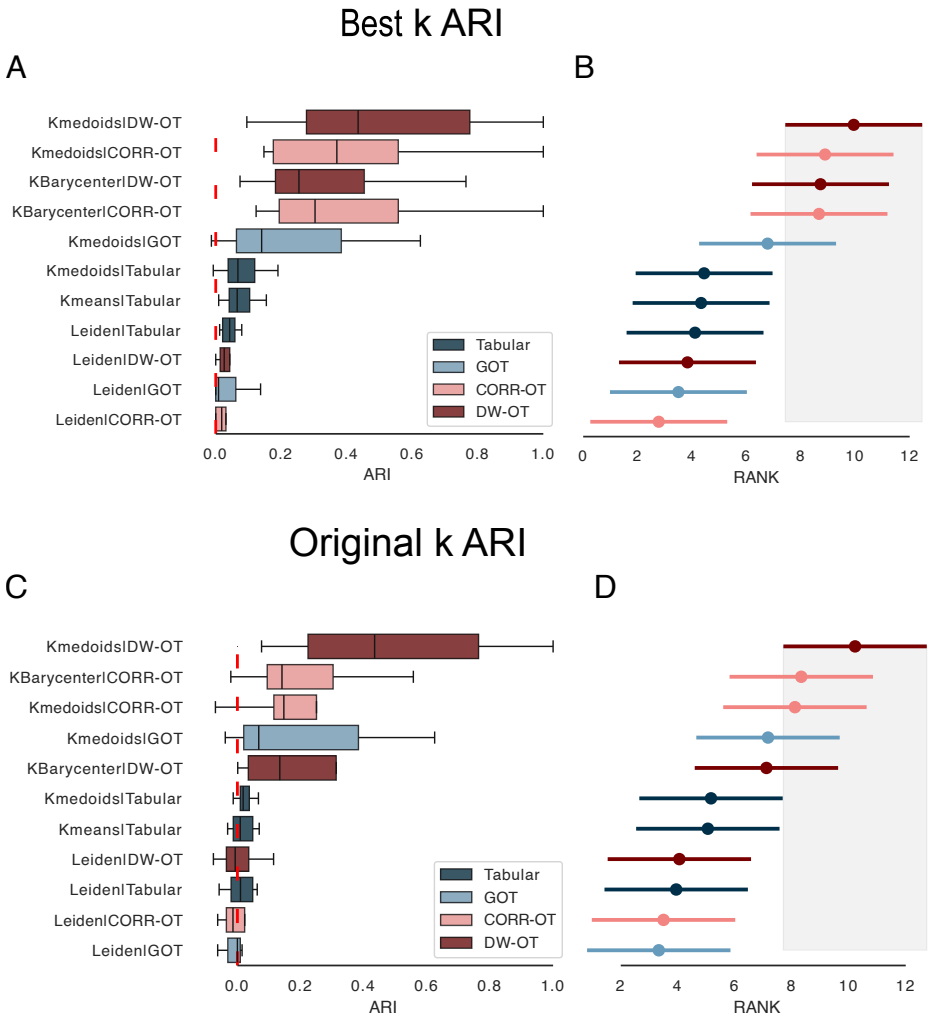


Figure 5.2: A–Boxplots indicate the maximum ARI value distribution (x-axis) for all evaluated methods over five scRNAseq data sets. B–Ranking values (mean and standard deviation) for each method and data set regarding the maximum ARI value. The highest ranking indicates the highest ARI. The gray area indicates the 95% confidence interval of the Friedman and Nemenyi posthoc-test. Methods whose average values are not within the gray area have significantly lower rankings than the top-ranked methods. For both A and B, methods are ranked by average in decreasing order. C and D are the same as A and B for the ARI estimates, with the number of clusters equal to the number of original labels.

Table 5.2: Adjusted Random Index values for all clustering methods and datasets. Best performing methods are highlighted in **bold**.

Method / Dataset	Kidney CKD/AKI	Breast Cancer	Pancreas Adenocarcinoma	COVID	Kidney Renal Carcinoma	Lung Adenocarcinoma	Myocardial Infarction
KBarycenter:CORR-OT	**0.303182**	0.094386	-0.020883	0.129534	**1.000000**	**0.183582**	0.141162
KBarycenter|DW-OT	0.134250	0.000999	0.034453	0.001123	**1.000000**	0.141784	0.036802
Kmeans|Tabular	0.015940	0.009027	0.069089	0.048706	-0.022337	-0.002409	-0.030983
Kmedoids|CORR-OT	0.087983	0.115361	-0.069666	0.147130	**1.000000**	0.134414	0.250494
Kmedoids|DW-OT	0.076468	**0.276956**	**1.000000**	0.223579	**1.000000**	0.148882	**0.713365**
Kmedoids|GOT	-0.038452	0.067632	0.109974	**0.413028**	-0.013245	0.048071	0.624783
Kmedoids|Tabular	0.018215	-0.007536	0.066567	0.037638	-0.013245	0.024293	0.009031
Leiden|CORR-OT	-0.014061	0.022934	-0.034972	0.023210	0.114583	-0.023132	-0.039031
Leiden|DW-OT	-0.017802	0.018484	-0.034972	0.035998	0.114583	-0.007003	0.165712
Leiden|GOT	-0.023747	0.000000	-0.030510	0.014790	-0.062500	0.000000	0.094465
Leiden|Tabular	-0.057877	0.018171	-0.046670	0.048582	0.060302	0.009701	0.062769

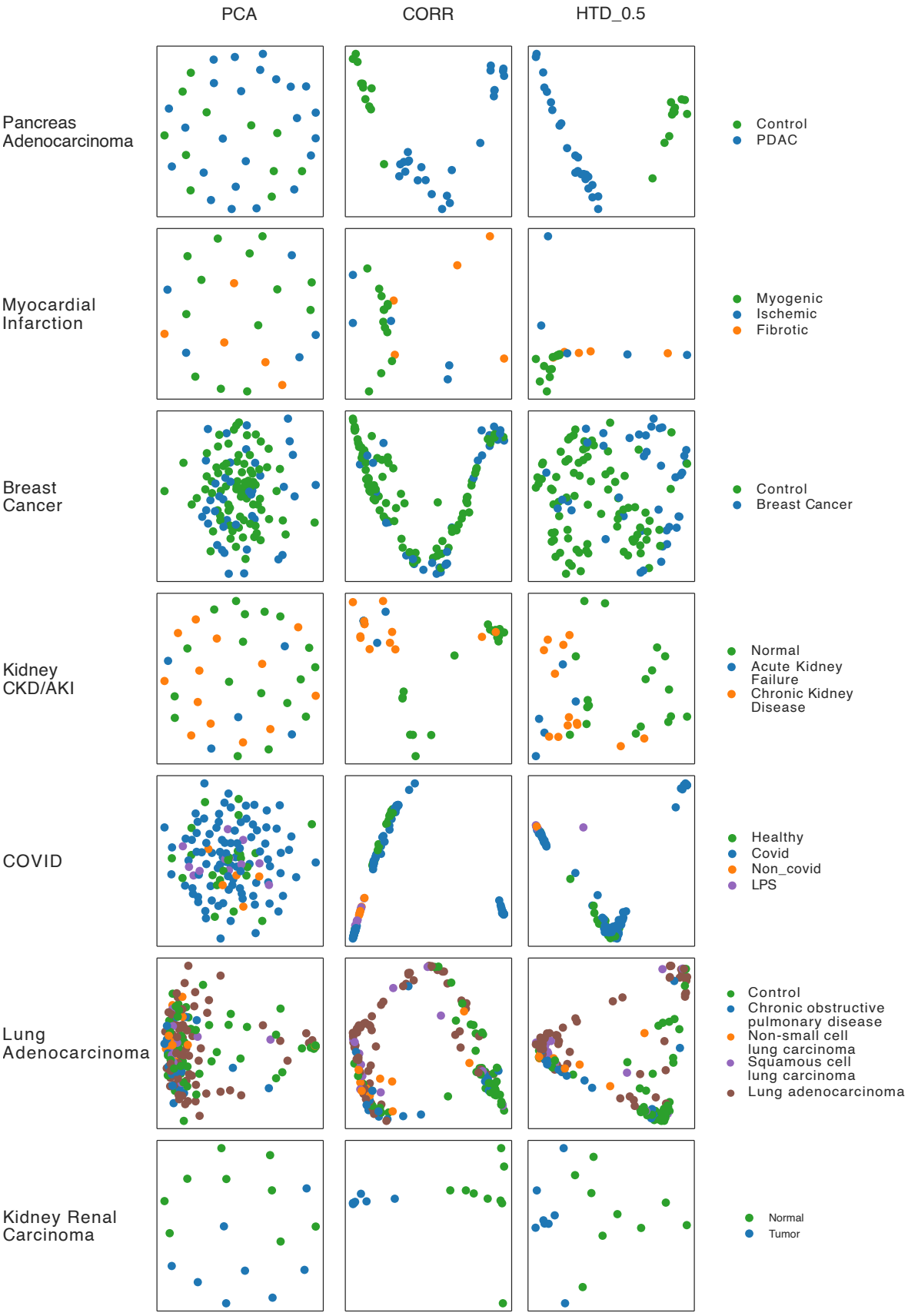


Figure 5.3: Patient CCC 2D Embeddings

PHATE two-dimensional embeddings of the distances matrices estimated by `Tabular`, `CORR-OT`, and `DW-OT` for all evaluated data sets. Colors correspond to the original labels.

The benchmark results are shown in Figure 5.2A-D, and Table 5.2. We observe that `DW-OT` with k-medoids has the highest mean ARI for the maximum ARI evaluation (Figure 5.2A). A Friedman-Nemenyi test indicates that `DW-OT` with *k*-medoids has the highest ranking and significantly outperforms `Tabular` based baseline approaches and the use of Leiden clustering (Figure 5.2B). `DW-OT` with *k*-medoids obtains the highest mean ARI for the number of clusters equal to the number of classes (Figure 5.2C-D). A Friedman–Nemenyi test indicates that this *k*-medoids variant also outperforms `Tabular` based baseline approaches and the use of Leiden clustering . Figure 5.3 shows embeddings (Moon et al., 2017) obtained from distances generated in this study. Regarding Rand Index (Figure A.2C), we observe similar rankings of methods and an average Rand index varying from 0.6 to 0.8 for DW-OT. These results underline the advantage of `DW-OT` with k-medoids, which is the only approach incorporating both directionality and connectivity of CCC graphs to cluster the samples.

5.4.2 Robustness analysis

Here, we describe how to assess different robustness aspects of scACCorDiON. As an initial assay, we analyzed the robustness of scACCorDiON with respect to cell-type annotation resolution. To this end, we evaluated clustering performance at two annotation levels (major cell types and refined cell types/sub-states) using the myocardial infarction (MI) and renal clear cell carcinoma (RCC) datasets. Moreover, the performance of under-sampling the number of cells in the PDAC dataset was also evaluated, i.e., we randomly removed up to 75% of cells and performed LR and scACCorDiON analysis. For each stratum, five random samples were generated.

Table 5.3: Results using two different cell type annotation granularity for myocardial Infarction (MI) and renal clear cell carcinoma (RCC). These results are based on *k-medoids* using the DW-OT distances. Best performing annotation are highlighted in **Bold**

	ARI	RI
RCC	**1.000000**	**1.000000**
RCC: Major Annotation	0.763889	0.882353
MI	**0.713365**	**0.861660**
MI: Major Annotation	0.434654	0.707510

Next, we evaluate whether the granularity of the cell type annotation impacts the overall results in the MI and RCC data sets, which provide two levels of cell type annotation. Results (Table 5.3) indicate this in both cases; DW-OT obtains the best results at coarse annotation levels. This supports that cellular sub-state annotation is beneficial for cell-cell communication prediction. In addition, we assess the robustness of our approach with respect to scRNA-seq quality, i.e., the recovery of cells per cluster. A cell down-sampling analysis in the PDAC data indicates that scACCorDiON only slightly deteriorates (0.12 ARI) when only 37.5% of the cells are kept.

One assumption of scACCorDiON is that masses are conserved, i.e., the same level of cell-cell communication is present in distinct samples. To evaluate this, we contrast the performance of the balanced and unbalanced OT formulations (See Chapter A). The comparison of DW-OT with the

balanced OT vs. unbalanced versions indicates no statistical difference between approaches. Nevertheless, the balanced DW-OT obtains the higher ARI (and RI) scores (Figure A.3). This supports the feasibility of the hypothesis of mass conservation as employed by scACCorDiON.

5.4.3 Case study: Pancreatic Ductal Adenocarcinoma

To evaluate the power of scACCorDiON in the detection of novel sub-clusters, we perform a Silhouette analysis (Rousseeuw, 1987) to identify data sets with a higher number of clusters than true labels (Figure A.4). Interestingly, we observe that for the pancreas adenocarcinoma (PDAC) data, scACCorDiON predicts a sub-cluster associated with controls and two sub-groups associated with PDAC samples (Figure 5.4A).

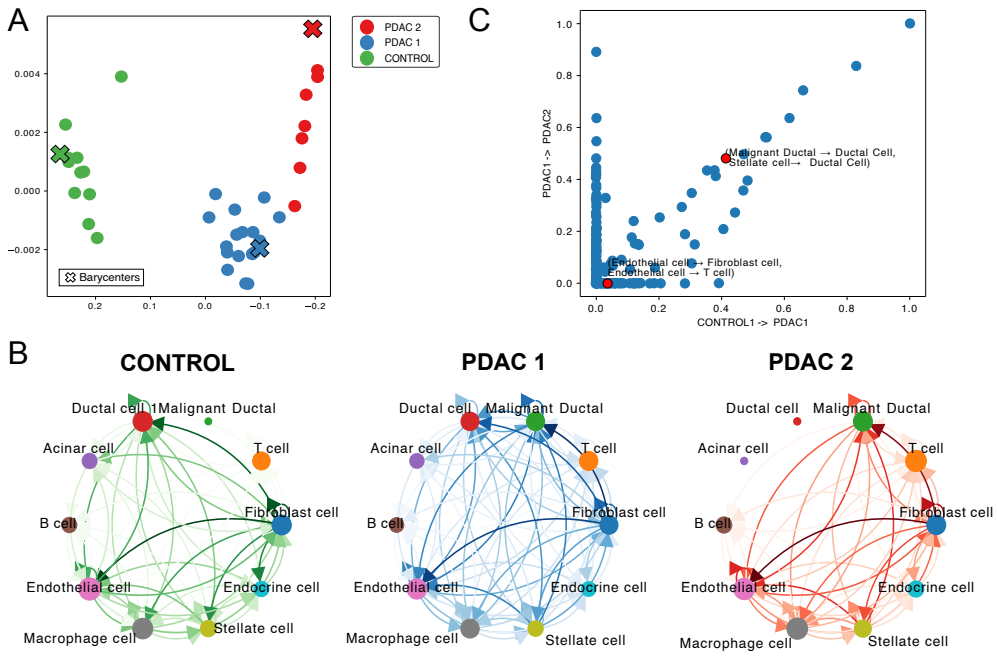


Figure 5.4: Sub-cluster analysis in pancreas adenocarcinoma A) MDS plot containing the clustering results on pancreas adenocarcinoma. One cluster contains all control samples (Control), and two clusters contain disease samples (PDAC 1 and 2). B) Directed CCC graphs of barycenters of the detected sub-clusters. Edge thickness indicates the strength of the cell-cell interactions. Node sizes are determined by node pagerank, and low-representation edges were filtered to improve visualization. C) Scatter plot with the transport map between the barycenter of the control and PDAC 1 (y-axis) and PDAC 1 and PDAC 2 clusters (y-axis). Every dot shows the mass transported between two cell-cell pairs in the comparisons.

As displayed in Figure 5.4B, PDAC 1 has overall increased communication, particularly interactions related to ductal, malignant ductal, and fibroblast cells. PDAC 2 demonstrates a loss of communication regarding Acinar and Ductal cells, while we observe an increase in communication related to Malignant Ductal, B, and Endocrine cells. The prominent signal related to the Malignant Ductal Cells indicates that PDAC 2 clusters are, possibly, linked to more advanced disease stages than the PDAC 1 cluster (Peng et al., 2019).

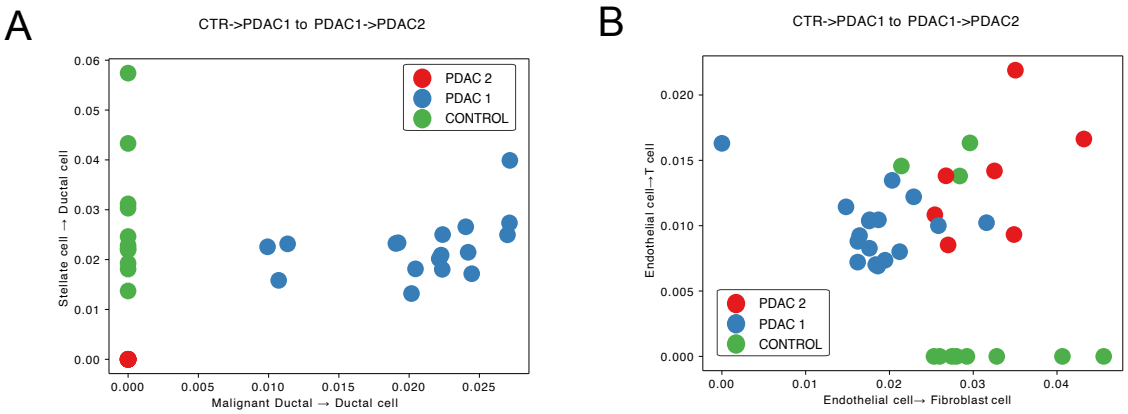


Figure 5.5: Sub-cluster analysis in pancreas adenocarcinoma A-B) Scatter plot with signals associated with selected cell-cell pairs for all the samples shown in Figure 5.4A). Colors correspond to the sample cluster.

To understand CCC events related to transitions from control to early disease (Control -> PDAC 1). Between mild and advanced disease (PDAC 1->PDAC 2), we contrast the transport maps (Γ) between the barycenters of these pairs of groups (Figure 5.4C). We observe that pairs of CCC interactions with high transport masses discriminate well the detected groups (Figure 5.5A-B).

We next make use of the LR analysis from CrossTalkeR (Nagai et al., 2021) to further investigate the interactions associated with the communication between Malignant Ductal cells and Ductal cells (Figure 5.6). We observed high expression in ERBB and EGFR receptors' interactions among the top LR pairs. These receptors were previously assigned to be related to pancreatic intraepithelial neoplasia (PanIN) (Ghasemi et al., 2014; Meyers et al., 2020), described as a precursor stage of Pancreas Adenocarcinoma. Moreover, ligands secreted by Malignant Ductal cells include matrix metalloproteinase-7 (MMP7) and galactins (LGALS-3/LGALS-3BP/LGALS-9), which have been recently shown to be expressed in malignant cells (Crawford et al., 2002). These results support an association between CCC changes in the tumor microenvironment and PDAC progression.

PDAC Ligand-Receptor survival analysis

Despite the number of CCC inference tools and efforts in the field, there are still no gold standards for validating them. To cover this gap and as a feature of scACCorDiON, Der. rer. nat. Tiago Maié and I propose validating CCC-related gene pairs using survival analysis, as in Figure 5.7. To this end, we use a Cox model to assess the predictive power of survival using the selected CCC pairs. This model assumes the hazard ratio does not depend on time. Let $h(t)$ be a function that measures hazard at a specific time t and $x_1, x_2, \cdots, x_p$ be the LR pairs. $h_0(t)$ is the baseline hazard when all explanatory variables are zero. The hazard model can be written as:

$$\ln \frac{h(t)}{h_0(t)} = b_1 x_1 + \cdots + b_p x_p \tag{5.11}$$

To validate the LR predicted in the previous PDAC analysis, we use the approach aforementioned with LR pairs and individual genes by using bulk RNA-seq data (n=177) from the PAAD TCGA data (Raphael et al., 2017). We considered the top 10 positive (PDAC 1) and negative (PDAC 2) hits

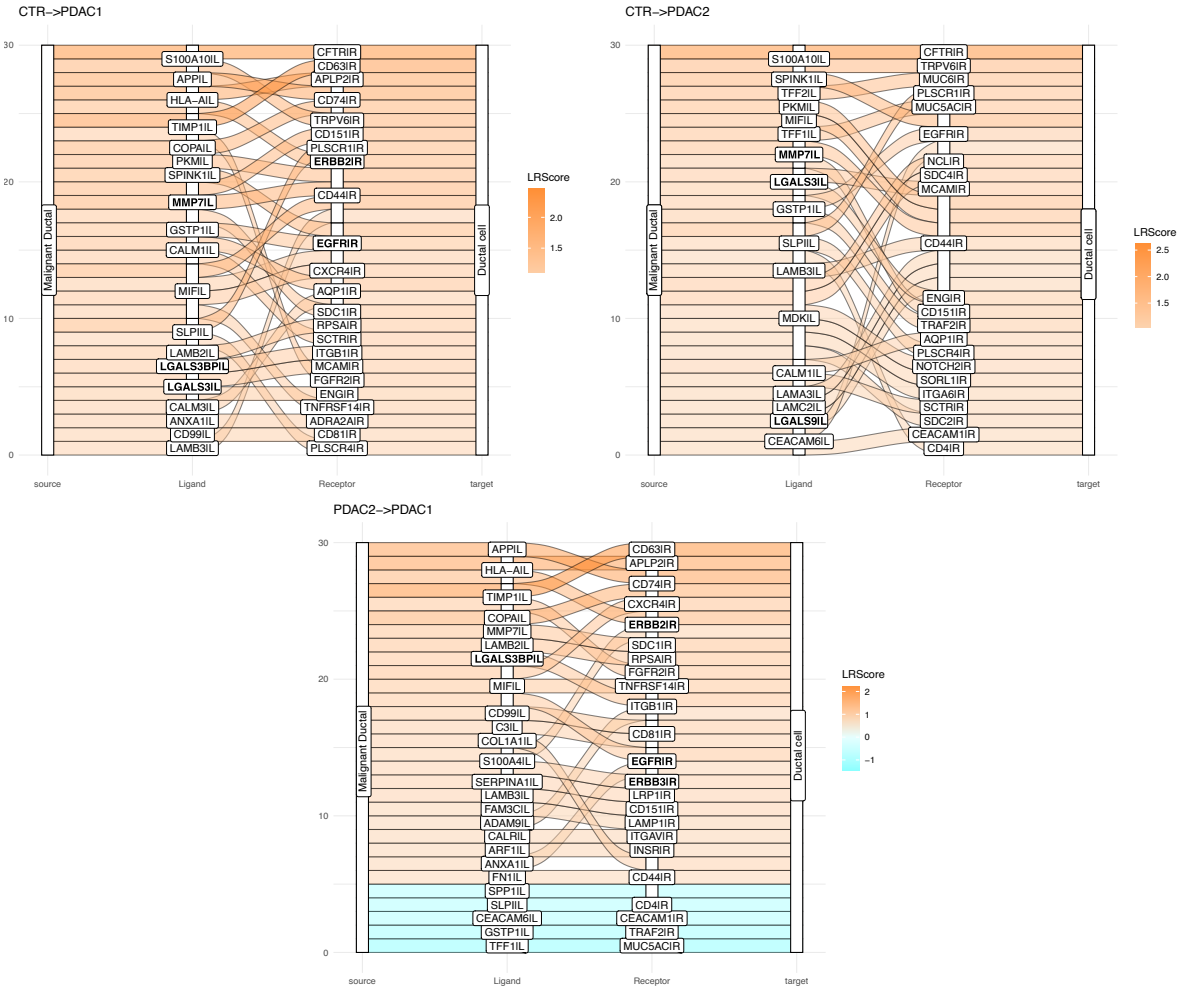


Figure 5.6: Top 20 ligand-receptor pairs (highest absolute ligand-receptor expression values) between malignant ductal cells towards ductal cells between all three group pairs. We highlight LR pairs discussed in the text.

for the cell-pair (Malignant Ductal Cells, Ductal Cell) and performed survival analysis using the Cox Proportional-Hazards model (Andersen and Gill, 1982), always keeping the stage information as a feature in each model.

For LR pairs with increased expression in PDAC 2 cells (vs. PDAC 1), we observed 3 LRs with significant survival associations, where a higher LRScore leads to a worse prognosis. Two receptors also showed a significant signal (Supp. Table 5.4). When comparing PDAC 1 with PDAC 2, we detected four significant LR interactions, four significant receptors, and two ligands (Supp. Table 5.5). Also, in both cases, LR expression pairs displayed the highest significance (top 2 for PDAC 2 and top 3 for PDAC 1), which suggests that the composition expression of the predicted LR pairs is a better predictor of survival than individual genes. One example of a pair whose significance is higher than its individual ligand and receptors is MMP7 -> SDC1 (Supp. Tab. 5.5). As previously mentioned, MMP7 has been associated with malignant ductal cells, and current research supports its role in early PDAC stages and tumor metastasis, leading to poorer prognosis (Van Doren, 2022). Altogether, this analysis supports the translational potential of predictions by scACCorDiON.

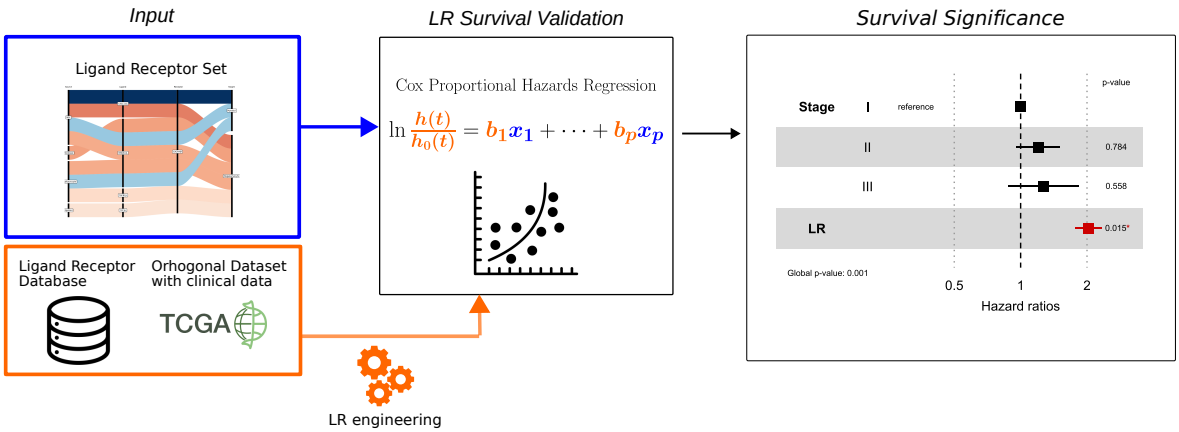


Figure 5.7: Schematic of the validation pipeline developed for assessing the survival prediction power of LR pairs. In the context of scACCorDiON, the LR set is selected by thresholding the highly expressed LR pairs, giving cell-cell interaction of interest.

Table 5.4: Survival Analysis hits using the top 10 Malignant Ductal->Ductal cell in PDAC 2
Cox Proportional Hazards models results using the top 10 Malignant Ductal->Ductal cell related to the interaction related to PDAC 2. (*) indicates statistically significant hits for survival prediction(p-value<0.05). Terms in **bold** indicate individual genes (ligand or receptor), which are also detected as significant as an LR pair.

	LR model			**Ligand model**			**Receptor model**		
Cluster	**Terms**	**OR**	**p-value**	**Terms**	**OR**	**p-value**	**Terms**	**OR**	**p-value**
PDAC 2	**SPP1_CD44**	1.547	0.004*	GRN	0.626	0.078	PLD2	0.458	0.017*
	GRN_SORT1	0.494	0.019*	GSTP1	1.441	0.169	**CD44**	1.467	0.024*
	GSTP1_TRAF2	1.851	0.037*	SPP1	1.092	0.203	MUC5AC	1.08	0.054
	ARF1_PLD2	0.458	0.106	CEACAM6	1.094	0.282	SORT1	0.77	0.202
	TFF1_MUC5AC	1.086	0.144	COL9A2	0.894	0.345	Stage II	1.619	0.28
	COL9A2_CD44	0.814	0.198	TFF2	0.957	0.484	CD4	0.936	0.593
	SLPI_CD4	0.826	0.301	MMP7	1.044	0.549	TRAF2	1.131	0.611
	TFF2_MUC6	0.954	0.35	Stage II	1.254	0.595	MUC6	1.021	0.613
	CEACAM6_CEACAM1	1.099	0.373	ARF1	0.887	0.771	Stage III+	1.135	0.863
	MMP7_ERBB4	0.958	0.52	TFF1	0.986	0.844	ERBB4	0.988	0.875
	Stage II	1.193	0.668	Stage III+	0.932	0.922	CEACAM1	1.015	0.895
	Stage III+	0.836	0.801	SLPI	1.007	0.945			

Comparison with Tensor-cell2cell

As stated above, only a few methods can perform sample-specific analysis in the CCC context. Here, we leverage the supervised tensor-based approach from (Armingol et al., 2022) for inspecting the subgroups revealed by scACCorDiON and the original sample labels. If we check the factors by comparing the two known class labels (Control vs. PDAC), we observe that two factors (6 and 8) are significantly associated with controls; 5 factors (2, 3, 4, 5, and 7) with PDAC and one factor (1) is not related to the known labels (Figure A.5A). Interestingly, by providing the clustering from scACCorDiON, we observe some factors to be related to PDAC 1 (Factors 4 and 5) and others to PDAC 2 expression (Factors 1 and 7) (Figure A.5B). These different loadings per sub-cluster support the biological relevance of these sample clusters.

Tensor-cell2cell can indicate cell-cell networks and LR pairs related to each factor for interpretation. Two factors (1 and 4) are related to the Malignant Ductal cell -> Ductal cell interaction (Fig-

Table 5.5: Survival Analysis hits using the top 10 Malignant Ductal->Ductal cell in PDAC 1 Cox Proportional Hazards models results using the top 10 Malignant Ductal->Ductal cell LR pairs related to PDAC 1. (*) indicates statistically significant hits for survival prediction(p-value<0.05) and Terms in **bold** indicate individual genes (ligand or receptor), which are also detected as significant as an LR pair.

	LR model			**Ligand model**			**Receptor model**		
Cluster	**Terms**	**OR**	**p-value**	**Terms**	**OR**	**p-value**	**Terms**	**OR**	**p-value**
PDAC 1	C3_CD81	1.978	0.004*	**MMP7**	1.169	0.026*	ERBB3	1.428	0.022*
	MMP7_SDC1	1.408	0.005*	**TIMP1**	0.65	0.026*	RPSA	1.665	0.022*
	CD99_CD81	0.427	0.017*	C3	1.199	0.105	ITGB1	1.518	0.023*
	TIMP1_FGFR2	0.705	0.038*	LAMB2	0.706	0.168	CD63	0.562	0.032*
	S100A4_ERBB3	1.211	0.123	S100A4	1.108	0.186	SDC1	1.196	0.144
	TIMP1_CD63	0.704	0.16	CD99	0.787	0.283	CD81	0.805	0.23
	COPA_CD74	0.629	0.321	COPA	0.893	0.743	FGFR2	0.886	0.231
	LAMB2_RPSA	0.754	0.429	LGALS3BP	0.944	0.788	CD74	1.165	0.302
	Stage III+	0.681	0.6	Stage II	1.083	0.86	Stage III+	0.675	0.595
	APP_CD74	0.816	0.652	APP	1.027	0.922	Stage II	1.213	0.656
	Stage II	0.868	0.747	Stage III+	0.995	0.995			
	LGALS3BP_ITGB1	0.988	0.969						

ure A.6A). Factor 1 is more prominent in PDAC 2 cells, and factor 4 is more prominent in PDAC 1 cells (Figure A.5). Pathway analysis with progeny (Schubert et al., 2018) indicates that these factors are mostly related to similar pathways, such as TNFa, NFKb, and EGFR (Figure A.6B). In contrast, the PDAC 1 factor had higher activity for Hypoxia and TGFb pathways. An equivalent analysis of Ductal cell -> Ductal cell interactions predictions from scACCorDiON also finds similar pathway pathways except for a lack of TGFb signal (A.6C).

We next performed survival analysis equivalent to the one above by selecting the top 10 ligand-receptor loadings for factor 1 and factor 4 (Table 5.6 and Table 5.7). To our surprise, only a single receptor pair and four genes were predictors of survival compared to seven pairs and ten genes in scACCorDiON/CrossTalkeR predictions. A possible reason for this is that Tensor-cell2cell factors are not specific to cell-cell pairs or a sample group (PDAC 1 or PDAC 2), as is the case for CrossTalkeR. Nevertheless, Tensor-cell2cell and CrossTalkeR can be seen as complementary and interpretable approaches, which can be used in a complementary manner.

5.5 Discussion

scRNAseq-based LR analysis enables the inference of CCC events related to complex diseases. However, sample-specific analysis, crucial for understanding CCC events in patient cohorts, has only been addressed to a limited extent so far. Here, we explore the problem of clustering samples that share similar cell communication patterns by modeling sample-specific CCC as directed and weighted graphs. We propose a graph-based optimal transport framework that finds optimal probabilistic mappings between cell communication signals and cell-cell graphs. Furthermore, this framework allows us to measure the distance between any two directed weighted graphs (regarding a Wasserstein distance) and estimate "average directed weighted graphs" (barycenters) representing typical CCC patterns within a group of samples. Our algorithm is currently unique in that it allows both computing dis-

Table 5.6: Survival Analysis hits using the top 10 Malignant Ductal->Ductal cell in Factor4 Cox Proportional Hazards models results using the top 10 LR pairs related to Factor 4(PDAC 1 high). (*) indicates statistically significant hits for survival prediction(p-value<0.05)

	LR model			Ligand model			Receptor model		
Cluster	**Terms**	**OR**	**p-value**	**Terms**	**OR**	**p-value**	**Terms**	**OR**	**p-value**
Factor 4 (PDAC 1)	S100A4_ERBB3	1.67	0.08	HSPG2	0.707	0.036*	ERBB3	1.497	0.017*
	HSPG2_SDC1	0.611	0.103	APOE	0.841	0.042*	KCNQ1	0.814	0.056
	APOE_LSR	0.778	0.117	CALM1	0.634	0.089	SDC1	1.274	0.065
	FN1_SDC1	1.704	0.194	FN1	1.241	0.166	EGFR	1.33	0.096
	S100A4_ERBB2	0.688	0.275	HMGB1	0.749	0.386	LSR	0.863	0.384
	CALM1_KCNQ1	0.837	0.306	S100A4	1.063	0.421	Stage II	1.258	0.611
	COL3A1_DDR1	1.288	0.324	COL3A1	1.174	0.573	ERBB2	1.013	0.919
	HMGB1_SDC1	1.354	0.413	Stage II	1.178	0.702	DDR1	0.994	0.975
	Stage II	1.303	0.538	COL1A2	0.962	0.891	Stage III+	0.991	0.99
	S100A4_EGFR	0.983	0.947	Stage III+	0.995	0.994			
	Stage III+	0.969	0.965						
	COL1A2_SDC1	0.983	0.966						

Table 5.7: Survival Analysis hits using the top 10 Malignant Ductal->Ductal cell in Factor1 Cox Proportional Hazards models results using the top 10 LR pairs related to Factor 1(PDAC 2 high). (*) indicates statistically significant hits for survival prediction (p-value<0.05)

	LR model			Ligand model			Receptor model		
Cluster	**Terms**	**OR**	**p-value**	**Terms**	**OR**	**p-value**	**Terms**	**OR**	**p-value**
Factor 1 (PDAC 2)	TFF3_CXCR4	0.819	0.031*	LAMC2	1.35	0.013*	PLSCR1	1.42	0.098
	CDH1_IGF1R	0.569	0.086	TFF3	0.953	0.329	CXCR4	0.856	0.192
	LAMC2_CD151	1.483	0.23	GDF15	1.055	0.463	CD44	1.262	0.192
	CDH1_PTPRF	1.36	0.279	Stage II	1.251	0.604	TGFBR2	1.213	0.342
	SLPI_PLSCR1	1.21	0.357	SPINT1	1.104	0.647	CD151	1.242	0.372
	SPINT1_ST14	0.798	0.483	CDH1	0.918	0.73	IGF1R	0.866	0.387
	LAMC2_CD44	1.204	0.53	MMP7	1.017	0.813	Stage II	1.302	0.546
	GDF15_TGFBR2	1.064	0.6	Stage III+	0.879	0.858	Stage III+	1.222	0.777
	Stage II	1.214	0.647	CGN	0.973	0.884	PTPRF	1.042	0.819
	MMP7_CD44	1.032	0.794	SLPI	1.012	0.899	ST14	1.005	0.978
	Stage III+	0.854	0.823						
	CGN_TGFBR2	0.945	0.835						

tances and clustering of directed weighted graphs. We have applied our DW-OT algorithm to calculate CCC graphs estimated in scRNAseq with large cohorts and found that it outperforms other algorithms. An interesting result is that clustering worked better at fine resolution for data sets where coarse- and fine-level annotations were provided. This supports the idea that cellular sub-states are important in understanding cell-cell communication mechanisms.

In the DW-OT method, the "mass" of cell-cell communication signals is conserved between patients, i.e., it assumes that the same amount of cell-cell communication is present in both normal and disease samples. By utilizing prior biological knowledge, other works have explored mass variations in optimization problems via unbalanced Optimal Transport (UOT) (Peyré et al., 2019), such as in cell development and proliferation (Schiebinger et al., 2019). There, prior knowledge is related to cell expansion or death, estimated from relevant pathway changes as supported by the expression profiles of the single cells. In our benchmarking, UOT did not improve our results. However, we lack a reliable source of prior biological knowledge on mass changes in cell-cell communication, as

ligand-receptor interactions follow complex Stoichiometric principles: increasing a ligand does not necessarily mean more signal to receptors due to saturation (Attie and Raines, 1995). Modeling mass changes in cell-cell communication is an interesting venue for further research.

We further showcased how both barycenter and transport matrices can be used to interpret communication events supporting detected clusters. This usage is exemplified in the pancreas adenocarcinoma data set, where DW-OT detected sub-clusters not characterized in the original study, presenting the data (Peng et al., 2019). Using the signatures of the identified groups, we conducted a survival-based analysis using the TCGA-PAAD dataset on top LR pairs. Interestingly, ligand-receptor expression was more significant than individual ligand and receptor genes. Moreover, we observed a tendency for higher enrichment in receptors than ligands, which potentially hints at the previously mentioned saturation aspect, i.e., an increase in ligands might not lead to an increase in cell-cell communication.

CHAPTER 6

Discussions and Future Work

This chapter presents an overview of the thesis's achievements and caveats. In this thesis, we motivate and introduce the comparative analysis of cell-cell communication graphs. Over the last few years, computational biology, particularly single-cell molecular data, has advanced in both data acquisition and analysis. In this chapter, we aim to provide an overview of the thesis's contributions and outline the limitations and future opportunities in CCC research.

6.1 Overall Discussion

With the ongoing development of cell-cell communication (CCC) research in computational biology, several interesting approaches and analyses have been presented over the last few years. Here, we aim to situate the representation of both CrossTalkeR and scACCorDiON in the field. Moreover, we discuss some limitations and the directions that we envision CCC research in the near future. Currently, more than a hundred tools (Armingol et al., 2024; Cesaro et al., 2025; Cheng et al., 2025) are available for performing various CCC-related tasks using transcriptomic data. Yet, in CCC research, a major challenge is the lack of ground-truth datasets for evaluating the existing tools (Armingol et al., 2024; Cesaro et al., 2025; Cheng et al., 2025).

6.1.1 Differential Graph Analysis

In the context of differential CCC analysis, CrossTalker is a pioneering approach that leverages the interaction nature of CCC via differential graph representations to identify condition-related interactions. To the best of our knowledge, our approach is the only method for using graph-topological measures to compare condition-based CCCs. At last, CrossTalker was successfully used by other groups and by us to investigate CCC schemes in complex diseases and/or tissues, such as bone marrow fibrosis (Leimkühler et al., 2021; Gleitz et al., 2025; Banjanin et al., 2026), atopic dermatitis (Sirvent et al., 2023), and aging (Iwahashi et al., 2024).

Besides the success of CrossTalkeR, some improvements are still to be addressed. In CrossTalker, we focus on intercellular communication, meaning that signalling cascades within cells are not considered. In technical terms, this results in unconnected components in the gene-cell networks. For better coverage of the full signalling cascade, methods that account for intracellular signalling are needed, which can be achieved by using an additional molecular data modality or by inferring gene regulatory programs from scRNAseq data.

Along with the rapid development of sequencing technologies, data throughput has increased massively. In terms of CCC, this increase can be directly mapped to an increase in the types of cells present in the dataset and the number of disease stages investigated simultaneously. On the CrossTalker side, the increase in the number of cell types causes a visualization issue, particularly in the circular layout used for CCC. Finding ways to visualize CCC and directed weighted graphs remains a computational open problem, and further investigation and benchmarking are ongoing. Furthermore, to address the need for more robust CCC analysis across multiple conditions, there is still a lack of approaches that simultaneously compare three or more disease stages. Notably, CrossTalkeR networks provide a data reservoir for implementing and developing node/edge metrics in the context of directed, weighted, and signed graphs.

6.1.2 Sample Level Graph Analysis

As mentioned above, CrossTalkeR only compares CCC graphs pairwise across disease stages. With advances in the amount of sequenced data and the availability of molecular data in large patient cohorts, studying CCC events in such cohorts can improve our understanding of early disease events and disease progression. To enable the identification of disease subgroups associated with CCC, we developed scACCorDiON. scACCorDiON leverages CCC-directed weighted graphs for a metric-learning and clustering framework. To the best of our knowledge, scACCorDiON is a pioneering unsupervised approach for analyzing CCC using scRNAseq cohort data. For metric learning and clustering, scACCorDiON is grounded in computational optimal transport. scACCorDiON enables metric learning across graphs with overlapping nodes. In PDAC, scACCorDiON recovers a disease subgroup and identifies key disease-related ligand-receptor interactions.

As discussed in Chapter 2, a gold standard for validating CCC inference methods remains elusive. Technologies such as spatial transcriptomics, spatial proteomics, and live-cell sequencing offer promising approaches to address this (Armingol et al., 2024; Cesaro et al., 2025; Cheng et al., 2025). Motivated by this, we leveraged rich clinical cancer data while developing scACCorDiON. We validated scACCorDiON predictions and corresponding ligand-receptor pairs using a survival approach. Validation occurs by measuring ligand-receptor pairs in selected source and target cells that discriminate disease subgroups in a dataset containing clinical data. Using the measurement and the clinical data, a regression model identifies pairs that are related to patient survivability. To our knowledge, we are the first to propose ligand-receptor-based survival analysis to validate CCC predictions.

To address the growing complexity of data, identifying uninformative cell-cell interactions while preserving the CCC graph's connectivity remains necessary. scACCorDiON would benefit from such methods, specifically, when computing the ground metric in our optimal transport approach. In addition to these considerations, a key challenge in sample-level analysis of scRNAseq data is the presence of batch effects and noise (i.e., variance from off-target covariates) (Joodaki et al., 2024). Thus, methods or regularizations that correct or reduce such effects would also improve scACCorDiON's results.

6.1.3 Beyond Cell-Cell Communication

The CCC representation adopted in both CrossTalkeR and scACCorDiON is a directed weighted graph (DWG). In both cases, the nodes of these graphs represent cell types, and ligand-receptor (LR) interactions govern their interactions. Each cell-cell interaction (CCI) is a summary of a set of ligand-receptor interactions, which possess two main measurements: gene expression and abundance. Such tools can also be useful in other real-world systems, such as ecological networks (studying the same species in different environments), social media networks (cross-social-network influence over time), and urban mobility networks (countries' connectedness over time). Following the translational aspects of CrossTalkeR and scACCorDiON, we provide the guidelines below for using both tools in the context of urban mobility networks.

Using scACCorDiON and CrossTalkeR in other contexts

In the case of CrossTalkeR, by assembling an interaction table as presented in Table 3.1, both single and comparative analysis can be performed. Rewriting such a table in a transport network context Table 6.1, we can represent `source` and `target` as nations and replace genes (i.e., ligand-receptor) with cities (i.e., city A and city B), each nation is related to a city with an airport.

Table 6.1: CrossTalkeR input example for traffic network

Source	Target	City A	City B	Cost
Germany	Netherlands	Aachen	Rotterdam	52
Germany	Germany	Aachen	Heidelberg	52
…	…	…	…	…
Denmark	Germany	Odense	Aachen	120
Germany	Spain	Aachen	Barcelona	80

Remark: To obtain meaningful results, we suggest that measurements are carefully examined before using CrossTalkeR. In the scRNAseq context, it's best practice to preprocess, normalize, and batch-correct datasets. In the traffic network above, suppose that the cost is obtained using different measurement machines; the batch to be corrected in this case is the sampling effect of the different equipment.

For scACCorDiON, users can use any directed, positive-weighted graph set that shares a node set. Note that for meaningful clustering of DWGs, a minimal set of graphs is needed; in scRNAseq, the datasets considered in our benchmark ranged from 10 to 40 cells (nodes).

Note that both tools are implemented and shared as open-source projects. Our aim is to enable reproducible, collaborative research in which methods implemented for the overall analysis can also be useful in other contexts and for researchers/students of diverse backgrounds.

6.2 Future Work

Research in CCC is ongoing; new technologies enable more precise, detailed data acquisition, leading to better predictions and a deeper understanding of molecular signal transduction (Cesaro et al., 2025;

Cheng et al., 2025). Recently, spatial transcriptomics (ST) measures location-dependent transcript detection. While initial approaches could measure many genes, measurements were obtained from fixed-size spots containing a mix of cells. On the other hand, novel technologies enable single-cell resolution measurements, but the number of genes measured per sequencing run is limited by predefined panels. In addition to the lower capture gene rate, ST requires robust cell segmentation, further complicating data analysis (Mitchel et al., 2026).

6.2.1 Spatial Transcriptomics and Cell-Cell Communication

Paracrine, justacrine, and autocrine are considered location-dependent CCC types (Armingol et al., 2020, 2024; Cesaro et al., 2025; Cheng et al., 2025). Building on this hypothesis and using ST sequencing, indirect benchmark approaches were developed by correlating the co-localization of cell-pair-based LR pairs with their expression. While this approach is a data-driven proxy for validating CCC, CCC complexity remains under-addressed in the context of spatial transcriptomics (Cesaro et al., 2025). However, the relationship between location and communication remains an open question. With new sequencing methods and biological assays, there is a hope for raising the standards for CCC prediction evaluation. And, broadening knowledge of tissue-, location-, and disease-dependent cell-cell interactions.

Novel ST approaches enable the measurement of transcripts at subcellular resolution (Mitchel et al., 2026). Moreover, cell segmentation data enables the incorporation of morphological features to better understand CCC behaviour. Altogether, integrating such information into current CCC tools appears to be a promising venue for single-cell-resolution CCC analysis.

On the computational side, there is a need for methods that scale with the amount of data generated by new sequencing methods. Current cell-cell communication models account for dozens of cell types. New sequencing methods that capture single-cell-level cell-cell communication increase the number of nodes in the networks to more than 100,000 cells per sample. To sum up, approaches that account for batch effects to improve inference are also needed.

6.2.2 Beyond Graphs and Multimodal Cell-Cell Communication

As aforementioned, the full CCC signaling landscape can be inferred from intracellular information. Currently, there are few methods for complementarily using multiple data modalities to characterize CCC schemes (Cesaro et al., 2025; Cheng et al., 2025). Although such approaches and methods are applied to multimodal data, they often yield conclusions rooted in a single modality; joint knowledge remains understudied. To enable joint-modality inference, complex data representations are needed.

Current approaches adopt a tensor representation (Armingol et al., 2025) to uncover the sample-size CCC schemes with protein measurements and metabolic profiles. Alternatively, representing the data in a high-order manner using concepts from persistent homology, hypergraphs (Hou et al., 2025), and multilayer graphs (Cheng et al., 2021; Trimbour et al., 2026) can be a powerful and promising venue for uncovering the causal and relational aspects of distinct molecular data types.

An example of a high-order signalling network is described in SigXTalk (Hou et al., 2025), where nodes are genes and hyperedges (i.e., interactions among more than two nodes) represent regulatory

relationships among them. Alternatively, a cell-cell communication-constrained triangulation approach could be implemented using ST spatial coordinates, representing the cell-cell communication simplicial complex on the set of single cells.

Altogether, the availability of novel high-throughput multimodal data opens an exciting avenue for more robust, engineered computational approaches. Allowing for the enhancement of understanding of novel and existing complex biological processes.

Bibliography

Agueh M. and Carlier G. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Andersen P. K. and Gill R. D. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

Aono A. H., Pimenta R. J. G., Garcia A. L. B., et al. The wild sugarcane and sorghum kinomes: Insights into expansion, diversification, and expression patterns. *Frontiers in Plant Science*, 12: 668623, 2021.

Armingol E., Officer A., Harismendy O., and Lewis N. E. Deciphering cell-cell interactions and communication from gene expression. *Nature reviews. Genetics*, page epub, nov 2020.

Armingol E., Baghdassarian H. M., Martino C., et al. Context-aware deconvolution of cell–cell communication with tensor-cell2cell. *Nature communications*, 13(1):3665, 2022.

Armingol E., Baghdassarian H. M., and Lewis N. E. The diversification of methods for studying cell–cell interactions and communication. *Nature Reviews Genetics*, pages 1–20, 2024.

Armingol E., Larsen R. O., Gale L., et al. Tensor-cell2cell v2 unravels coordinated dynamics of protein-and metabolite-mediated cell-cell communication. *BioRxiv*, pages 2022–11, 2025.

Arthur D., Vassilvitskii S., and others . k-means++: The advantages of careful seeding. In *Soda*, volume 7, pages 1027–1035, 2007.

Attie A. D. and Raines R. T. Analysis of receptor-ligand interactions. *Journal of chemical education*, 72(2):119, 1995.

Banjanin B., Nagai J., Mun Y., et al. Wnt-dependent spatiotemporal reprogramming of bone marrow niches drives fibrosis. *HemaSphere*, 10(2):e70309, 2026.

Bianconi E., Piovesan A., Facchin F., et al. An estimation of the number of cells in the human body. *Annals of human biology*, 40(6):463–471, 2013.

Bouland G. A., Mahfouz A., and Reinders M. J. Consequences and opportunities arising due to sparser single-cell rna-seq datasets. *Genome biology*, 24(1):86, 2023.

Boyd Z. M., Fraiman N., Marzuola J., et al. A metric on directed graphs and markov chains based on hitting probabilities. *SIAM Journal on Mathematics of Data Science*, 3(2):467–493, 2021.

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

Browaeys R., Gilis J., Sang-Aram C., et al. Multinichenet: a flexible framework for differential cell-cell communication analysis from multi-sample multi-condition single-cell transcriptomics data. *bioRxiv*, pages 2023–06, 2023.

Cesaro G., Nagai J. S., Gnoato N., et al. Advances and challenges in cell–cell communication inference: a comprehensive review of tools, resources, and future directions. *Briefings in Bioinformatics*, 26(3):bbaf280, 2025.

Cheng J., Zhang J., Wu Z., and Sun X. 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.

Cheng M., Jansen J., Reimer K., et al. Phlower-single cell trajectory analysis using decomposition of the hodge laplacian. *bioRxiv*, pages 2024–10, 2024.

Cheng X., Huang H., Su Y., et al. Cell-cell communication inference and analysis: biological mechanisms, computational approaches, and future opportunities. *arXiv preprint arXiv:2512.03497*, 2025.

Chowdhury S. and Mémoli F. The gromov–wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019.

Chowdhury S. and Needham T. Generalized spectral clustering via gromov-wasserstein learning. In *International Conference on Artificial Intelligence and Statistics*, pages 712–720. PMLR, 2021.

Crawford H. C., Scoggins C. R., Washington M. K., Matrisian L. M., and Leach S. D. Matrix metalloproteinase-7 is expressed by pancreatic cancer precursors and regulates acinar-to-ductal metaplasia in exocrine pancreas. *J. Clin. Invest.*, 109(11):1437–1444, June 2002.

Cuturi M. and Doucet A. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

CZI Single-Cell Biology S., Aevermann B., Assis P., et al. Cz cellxgene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *bioRxiv*, pages 2023–10, 2023.

Demetci P., Santorella R., Sandstede B., Noble W. S., and Singh R. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *BioRxiv*, 2020.

Dempster A. P., Laird N. M., and Rubin D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, 1977.

Demšar J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

Deng Z., Fan T., Xiao C., et al. Tgf-β signaling in health, disease and therapeutics. *Signal transduction and targeted therapy*, 9(1):61, 2024.

Diaz-Diaz F., Candellone E., González-Casado M. A., et al. Signed networks: theory, methods, and applications. *arXiv preprint arXiv:2511.17247*, 2025.

Dimitrov D., Türei D., Garrido-Rodriguez M., et al. Comparison of methods and resources for cell-cell communication inference from single-cell rna-seq data. *Nature Communications*, 13(1):3224, 2022.

Dong X., Thanou D., Frossard P., and Vandergheynst P. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23):6160–6173, 2016.

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

Euler L. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pages 128–140, 1741.

Fischer A., Suen C. Y., Frinken V., Riesen K., and Bunke H. Approximation of graph edit distance based on hausdorff matching. *Pattern Recognition*, 48(2):331–343, 2015.

Flores R. O. R., Lanzer J. D., Dimitrov D., Velten B., and Saez-Rodriguez J. Multicellular factor analysis of single-cell data for a tissue-centric understanding of disease. *eLife*, 12:e93161, 2023.

Forman R. Determinants of laplacians on graphs. *Topology*, 32(1):35–46, 1993.

Freeman L. C. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

Furutani S., Shibahara T., Akiyama M., Hato K., and Aida M. Graph signal processing for directed graphs based on the hermitian laplacian. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 447–463. Springer, 2019.

Geisler S., Li Y., Mankowitz D. J., et al. Transformers meet directed graphs. In *International conference on machine learning*, pages 11144–11172. PMLR, 2023.

Ghasemi R., Rapposelli I., Capone E., et al. Dual targeting of erbb-2/erbb-3 results in enhanced antitumor activity in preclinical models of pancreatic cancer. *Oncogenesis*, 3(8):e117–e117, 2014.

Ghosh A., Boyd S., and Saberi A. Minimizing effective resistance of a graph. *SIAM review*, 50(1): 37–66, 2008.

Gleich D. F. Pagerank beyond the web. *SIAM Review*, 57(3):321–363, 2015.

Gleitz H. F., Kramann R., and Schneider R. K. Understanding deregulated cellular and molecular dynamics in the haematopoietic stem cell niche to develop novel therapeutics for bone marrow fibrosis. *The Journal of Pathology*, 245(2):138–146, 2018.

Gleitz H. F., Dugourd A. J., Leimkühler N. B., et al. Increased cxcl4 expression in hematopoietic cells links inflammation and progression of bone marrow fibrosis in mpn. *Blood, The Journal of the American Society of Hematology*, 136(18):2051–2064, 2020.

Gleitz H. F., Fuchs S. N., Snoeren I. A., et al. Inhibiting the alarmin-driven hematopoiesis-stromal cell crosstalk in primary myelofibrosis ameliorates bone marrow fibrosis. *HemaSphere*, 9(8):e70179, 2025.

Grohe M. and Schweitzer P. The graph isomorphism problem. *Communications of the ACM*, 63(11): 128–134, 2020.

Hafemeister C. and Satija R. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1):296, 2019.

Hart P. E., Nilsson N. J., and Raphael B. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

Hou J., Zhao W., and Nie Q. Dissecting crosstalk induced by cell-cell communication using single-cell transcriptomic data. *Nature Communications*, 16(1):5970, 2025.

Hou R., Denisenko E., Ong H. T., Ramilowski J. A., and Forrest A. R. Predicting cell-to-cell communication networks using natmi. *Nature communications*, 11(1):5011, 2020.

Hubert L. and Arabie P. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

Isaacs J. T., Pili R., Qian D. Z., et al. Identification of abr-215050 as lead second generation quinoline-3-carboxamide anti-angiogenic agent for the treatment of prostate cancer. *The Prostate*, 66(16): 1768–1778, 2006.

Ishimoto C. K., Aono A. H., Nagai J. S., et al. Microbial co-occurrence network and its key microorganisms in soil with permanent application of composted tannery sludge. *Science of the Total Environment*, 789:147945, 2021.

Iwahashi N., Umakoshi H., Fujita M., et al. Single-cell and spatial transcriptomics analysis of human adrenal aging. *Molecular Metabolism*, 84:101954, 2024.

Jadbabaie A., Motee N., and Barahona M. On the stability of the kuramoto model of coupled nonlinear oscillators. In *Proceedings of the 2004 American Control Conference*, volume 5, pages 4296–4301. IEEE, 2004.

Jansen J., Reimer K. C., Nagai J. S., et al. Sars-cov-2 infects the human kidney and drives fibrosis in kidney organoids. *Cell Stem Cell*, 29(2):217–231, 2022.

Jia L., Gaüzère B., and Honeine P. Graph kernels based on linear patterns: theoretical and experimental comparisons. *Expert Systems with Applications*, 189:116095, 2022.

Jin S., Guerrero-Juarez C. F., Zhang L., et al. Inference and analysis of cell-cell communication using cellchat. *Nature communications*, 12(1):1088, 2021.

Joodaki M., Shaigan M., Parra V., et al. Detection of patient-level distances from single cell genomics and pathomics data with optimal transport (pilot). *Molecular systems biology*, 20(2):57–74, 2024.

Kanehisa M. The kegg database. In *'In silico' simulation of biological processes: Novartis Foundation Symposium 247*, volume 247, pages 91–103. Wiley Online Library, 2002.

Kaufmann L. and Rousseeuw P. Clustering by means of medoids. *Data Analysis based on the L1-Norm and Related Methods*, pages 405–416, 01 1987.

Kim N. and Yoon Y. Effective urban region representation learning using heterogeneous urban graph attention network (hugat). *IEEE Access*, 13:102602–102612, 2025.

Kumar T., Nee K., Wei R., et al. A spatially resolved single cell genomic atlas of the adult human breast. *bioRxiv*, pages 2023–04, 2023.

Kuppe C., Ramirez Flores R. O., Li Z., et al. Spatial multi-omic map of human myocardial infarction. *Nature*, 608(7924):766–777, 2022.

Lähnemann D., Köster J., and Szczurek E. o. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

Lake B. B., Menon R., Winfree S., et al. An atlas of healthy and injured cell states and niches in the human kidney. *Nature*, 619(7970):585–594, 2023.

Landherr A., Friedl B., and Heidemann J. A critical review of centrality measures in social networks. *Business & Information Systems Engineering*, 2:371–385, 2010.

Leimkühler N. B., Gleitz H. F., Ronghui L., et al. Heterogeneous bone-marrow stromal progenitors drive myelofibrosis via a druggable alarmin axis. *Cell Stem Cell*, 28(4):637–652, 2021.

Li B., Shi Y., Yu Q., and Wang J. Unsupervised cross-domain image retrieval via prototypical optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38,4, pages 3009–3017, 2024.

Liu Y., Wang C., Wang Z., et al. High-parameter spatial multi-omics through histology-anchored integration. *bioRxiv*, pages 2025–02, 2025.

Lloyd S. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Luecken M. D., Gigante S., Burkhardt D. B., et al. Defining and benchmarking open problems in single-cell analysis. *Research Square*, 2024.

Malod-Dognin N. and Pržulj N. L-graal: Lagrangian graphlet-based network aligner. *Bioinformatics*, 31(13):2182–2189, 2015.

Maretic H. P., El Gheche M., Chierchia G., and Frossard P. Got: an optimal transport framework for graph comparison. *Advances in Neural Information Processing Systems*, 32, 2019.

Maretic H. P., Gheche M. E., Chierchia G., and Frossard P. fgot: Graph distances based on filters and optimal transport. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7710–7718, Jun. 2022.

Meyers N., Gérard C., Lemaigre F. P., and Jacquemin P. Differential impact of the erbb receptors egfr and erbb2 on the initiation of precursor lesions of pancreatic ductal adenocarcinoma. *Scientific reports*, 10(1):5241, 2020.

Mitchel J., Gao T., Petukhov V., Cole E., and Kharchenko P. V. Impact and correction of segmentation errors in spatial transcriptomics. *Nature Genetics*, pages 1–11, 2026.

Monge G. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

Moon K. R., Dijk D.v, Wang Z., et al. Phate: a dimensionality reduction method for visualizing trajectory structures in high-dimensional biological data. *BioRxiv*, 120378, 2017.

Nagai J. S., Leimkühler N. B., Schaub M. T., Schneider R. K., and Costa I. G. CrossTalkeR: analysis and visualization of ligand–receptorne tworks. *Bioinformatics*, 37(22):4263–4265, 05 2021.

Nagai J. S., Costa I. G., and Schaub M. T. Optimal transport distances for directed, weighted graphs: a case study with cell-cell communication networks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9856–9860. IEEE, 2024.

Nagai J. S., Maié T., Schaub M. T., and Costa I. G. scaccordion: a clustering approach for explainable patient level cell–cell communication graph analysis. *Bioinformatics*, 41(5):btaf288, 2025.

Nemenyi P. B. *Distribution-free multiple comparisons.* Princeton University, 1963.

Nikolentzos G., Meladianos P., and Vazirgiannis M. Matching node embeddings for graph similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31,1, 2017.

Page L., Brin S., Motwani R., and Winograd T. The pagerank citation ranking: Bring order to the web. Technical report, Technical report, Stanford University, 1998.

Peng B., Zhu Y., Liu Y., et al. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*, 2024.

Peng J., Sun B.-F., Chen C.-Y., et al. Single-cell rna-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell research*, 29(9):725–738, 2019.

Peyré G., Cuturi M., and others . Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Randić M. and Klein D. Resistance distance. *J. Math. Chem*, 12:81–95, 1993.

Raphael B. J., Hruban R. H., Aguirre A. J., et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell*, 32(2):185–203, 2017.

Razmkhah F., Kim S., Lim S., Dania A.-J., and Choi J. S100a8 and s100a9 in hematologic malignancies: from development to therapy. *International journal of molecular sciences*, 24(17):13382, 2023.

Rieder F., Nagy L. E., Maher T. M., et al. Fibrosis: cross-organ biology and pathways to development of innovative drugs. *Nature Reviews Drug Discovery*, pages 1–27, 2025.

Riesen K. and Bunke H. Approximate graph edit distance computation by means of bipartite graph matching. *Image and Vision computing*, 27(7):950–959, 2009.

Robinson M. D., McCarthy D. J., and Smyth G. K. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.

Rosvall M., Axelsson D., and Bergstrom C. T. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.

Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Schieber T. A., Carpi L., Díaz-Guilera A., et al. Quantification of network structural dissimilarities. *Nature communications*, 8(1):13928, 2017.

Schiebinger G., Shu J., Tabaka M., et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Schneider R. K., Mullally A., Dugourd A., et al. Gli1+ mesenchymal stromal cells are a key driver of bone marrow fibrosis and an important cellular therapeutic target. *Cell stem cell*, 20(6):785–800, 2017.

Schraivogel D., Gschwind A. R., Milbank J. H., et al. Targeted perturb-seq enables genome-scale genetic screens in single cells. *Nature methods*, 17(6):629–635, 2020.

Schubert E. and Rousseeuw P. J. Faster k-medoids clustering: improving the pam, clara, and clarans algorithms. In *Similarity Search and Applications: 12th International Conference, SISAP 2019, Newark, NJ, USA, October 2–4, 2019, Proceedings 12*, pages 171–187. Springer, 2019.

Schubert E. and Rousseeuw P. J. Fast and eager k-medoids clustering: O (k) runtime improvement of the pam, clara, and clarans algorithms. *Information Systems*, 101:101804, 2021.

Schubert M., Klinger B., Klünemann M., et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature communications*, 9(1):20, 2018.

Shaw M. E. Some effects of unequal distribution of information upon group performance in various communication nets. *The journal of abnormal and social psychology*, 49(4p1):547, 1954.

Shubin M. A. Discrete magnetic laplacian. *Communications in mathematical physics*, 164(2):259–275, 1994.

Sikkema L., Ramírez-Suástegui C., Strobl D. C., et al. An integrated cell atlas of the lung in health and disease. *Nature Medicine*, pages 1–15, 2023.

Silva M. E., Gaunt R. E., Ospina-Forero L., Jay C., and House T. Comparing directed networks via denoising graphlet distributions. *Journal of Complex Networks*, 11(2):cnad006, 2023.

Sirvent S., Vallejo A. F., Corden E., et al. Impaired expression of metallothioneins contributes to allergen-induced inflammation in patients with atopic dermatitis. *Nature Communications*, 14(1): 2880, 2023.

Stephenson E., Reynolds G., Botting R. A., et al. Single-cell multi-omics analysis of the immune response in covid-19. *Nature medicine*, 27(5):904–916, 2021.

Stuart T., Butler A., Hoffman P., et al. Comprehensive integration of single-cell data. *cell*, 177(7): 1888–1902, 2019.

Takatsu A. On wasserstein geometry of gaussian measures. In *Probabilistic approach to geometry*, pages 463–472. Mathematical Society of Japan, 2010.

Tamaru A., Hara J., Higashi H., Tanaka Y., and Ortega A. Optimizing k in knn graphs with graph learning perspective. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9441–9445. IEEE, 2024.

Teodorescu P., Pasca S., Jurj A., et al. Transforming growth factor β-mediated micromechanics modulates disease progression in primary myelofibrosis. *Journal of Cellular and Molecular Medicine*, 24(19):11100–11110, 2020.

Togninalli M., Ghisu E., Llinares-López F., Rieck B., and Borgwardt K. Wasserstein weisfeiler-lehman graph kernels. *Advances in neural information processing systems*, 32, 2019.

Traag V. A., Waltman L., and Van Eck N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

Trimbour R., Ramirez Flores R. O., Saez-Rodriguez J., and Cantini L. Modelling multicellular coordination by bridging cell-cell communication and intracellular regulation through multilayer networks. *bioRxiv*, pages 2026–01, 2026.

Van Doren S. R. Mmp-7 marks severe pancreatic cancer and alters tumor cell signaling by proteolytic release of ectodomains. *Biochemical Society Transactions*, 50(2):839–851, 2022.

Venkat A., Chew J., Rodriguez F. C., et al. Directed scattering for knowledge graph-based cellular signaling analysis. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9761–9765. IEEE, 2024.

Vishwanathan S. V. N., Schraudolph N. N., Kondor R., and Borgwardt K. M. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.

Wolf F. A., Angerer P., and Theis F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

Xu H., Luo D., Zha H., and Duke L. C. Gromov-wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pages 6932–6941. PMLR, 2019.

Xu M. and Chang L. Graph edit distance estimation: A new heuristic and a holistic evaluation of learning-based methods. *Proceedings of the ACM on Management of Data*, 3(3):1–24, 2025.

Younesi F. S., Miller A. E., Barker T. H., Rossi F. M., and Hinz B. Fibroblast and myofibroblast activation in normal tissue repair and fibrosis. *Nature Reviews Molecular Cell Biology*, 25(8): 617–638, 2024.

Young G. F., Scardovi L., and Leonard N. E. A new notion of effective resistance for directed graphs—part i: Definition and properties. *IEEE Transactions on Automatic Control*, 61(7):1727–1736, 2015a.

Young G. F., Scardovi L., and Leonard N. E. A new notion of effective resistance for directed graphs—part ii: Computing resistances. *IEEE Transactions on Automatic Control*, 61(7):1737–1752, 2015b.

Yu G. Thirteen years of clusterprofiler. *The Innovation*, 5(6), 2024.

Zhang Y., Parmigiani G., and Johnson W. E. Combat-seq: batch effect adjustment for rna-seq count data. *NAR genomics and bioinformatics*, 2(3):lqaa078, 2020.

Zvirblyte J., Nainys J., Juzenas S., et al. Single-cell transcriptional profiling of clear cell renal cell carcinoma reveals a tumor-associated endothelial tip cell phenotype. *Communications Biology*, 7 (1):780, 2024.

APPENDIX A

Appendix

A.1 MultiNicheNet

Altogather MultiNicheNet works in four main steps and requires as input a cohort scRNAseq dataset, a list of conditions (i.e, contrasts) to be compared, a list of cell-type pairs (i.e., list of source and target) of interest, the LR prior and the corresponding ligand-target matrix.

Step1: Differential Expression analysis

As an initial step, MultiNicheNet uses the scRNAseq raw count matrix to generate a pseudobulk count matrix for each cell type. In other words, the raw count matrix is summarized by summing the counts across cell types. Next, using the list of conditions to compare and the available batch information (i.e., unrelated covariates), a differential expression (DE) analysis is performed with edgeR (Robinson et al., 2010). As a result, the DE analysis results are stored for each gene, each cell type, and each contrast combination. Results include the gene fold change (i.e., the value indicating the increase or decrease in expression) and the significance from the DE statistical test.

Step2: Computing the normalized pseudobulk

After performing the DE analysis, MultiNicheNet computes the normalized pseudobulk expression matrix. To begin this step, Browaeys et al. (2023) normalized the original pseudobulk counts using the effective library size, as in edgeR (Robinson et al., 2010). Next, the normalized counts are scaled by 1,000,000 and then log2-transformed. To allow a zero-count value, a pseudocount of one is added to the normalized counts matrix prior to the log2 transformation. The log-normalized expression matrix is stored for each sample-cell type combination. Of Note, MultiNicheNet also provides the option for performing a batch correction step; to this end, the original count matrix obtained in Step 1 is corrected using the Combat-seq (Zhang et al., 2020) tool.

Step3: Ligand activity analysis and ligand-target inferece using NicheNet

Following the pipeline, MultiNicheNet subsequently uses the DE analysis results from Step 1. For each target cell-type and the contrast of interest, a set of upregulated (i.e., positive fold change) and downregulated (i.e., negative fold change) genes is determined. For selecting the genesets, Browaeys et al. (2023) selects significant genes (i.e., p-value < 0.05), and the log fold change thresholds are 0.5 and -0.5 for upregulated and downregulated genes, respectively.

After building the genesets, they are used to compute ligand activity using NicheNet (Browaeys et al., 2020). In this step, the AUPRC in target gene prediction is used as the ligand activity metric; all genes in the NicheNet ligand-target matrix serve as background for the prediction task.

Finally, target gene inference is performed per LR and contrast combination as in (Browaeys et al., 2020). This step only considers target gene predictions where the target gene is on the top-n (n=250) targets of the given ligand regarding the ligand-target model of the regulatory potential in NicheNet (Browaeys et al., 2020). The genesets, the ligand activities for each LR and contrast combinations, and the target gene inference results are stored in this step.

Step4: Identifying differential LR pairs

To identify differential LR pairs, five criteria are used. Each criterion is scaled to 01 and used in a weighted aggregation approach. We next list the criteria used to select differential LR pairs. The two

initial criteria relate to the DE LR within the respective source and target cells across all contrasts. To this end, for each contrast, the fold-change statistic for the ligand/receptor in the source/target is used to rank the genes. After this, the ranks are combined considering all contrast combinations. Finally, the aggregated rank is divided by the maximum rank possible, meaning that the maximal score of a ligand/receptor is 1. In addition to that, a p-value transformed score is computed by using a -log10 transformation of the p-value multiplied by the sign of the fold change statistic.

The following criterion is related to the ligand activity of the ligand within a target cell type for a specific comparison. The ligand activity values are normalized using the z-score and scaled with a min-max scaling. As the final ligand activity score, the maximum of the scaled-up activity score is selected. To reduce the impact of outliers, the upper and lower 1% quantiles of the activity scores are filtered out.

While the criteria presented above are obtained on a contrast basis, the following criteria consider the biological conditions instead. For the following criterion, the cell type expression matrix is averaged per condition, the min-max scaling is performed for each ligand per source and condition combinations. As a result of this process, this score captures the specificity of the ligand regarding conditions and source cells. For receptors, the score is computed in the same manner.

The last criterion represents the proportion of cells per condition in which the ligand/receptor is expressed desirably. LR pairs that are described in at least 5% of source/target cells are considered. To compute the final scores, the fractions of samples with sufficient expression are calculated per condition. Finally, the criteria are combined by a weight aggregation; the initial two criteria have a weight of 1, and the latest 3 have a weight of 2.

A.2 Unbalanced Optimal Transport

The *unbalanced* optimal transport map $\Gamma^* \in \mathbb{R}^{E\times E}$ for two probability distributions defined on the nodes of the line graph as induced by the two CCC graphs $\mathcal{G}^k$ and $\mathcal{G}^l$ can now be computed as:

$$\Gamma^* = \arg\min_{\Gamma} \langle \Gamma, C\rangle_F + \epsilon \cdot \mathrm{KL}(\Gamma, C) + \lambda_1 \cdot \mathrm{KL}(\Gamma\mathbf{1}||\mathbb{P}_{:,k}) + \lambda_2 \cdot \mathrm{KL}(\Gamma^T\mathbf{1}||\mathbb{P}_{:,l}) \tag{A.1}$$

$$\text{where} \quad \mathcal{S} = \left\{\Gamma \mid \Gamma\mathbf{1} = a, \Gamma^T\mathbf{1} = b, \Gamma_{ij} \geq 0\right\},$$

and the associated (induced) Wasserstein distance between the two CCC samples is:

$$d_W(\mathcal{G}^k, \mathcal{G}^l) = \arg\min_{\Gamma} \langle \Gamma, C\rangle_F + \epsilon \cdot \mathrm{KL}(\Gamma, C) + \lambda_1 \cdot \mathrm{KL}(\Gamma\mathbf{1}||a) + \lambda_2 \cdot \mathrm{KL}(\Gamma^T\mathbf{1}||b) = \langle \Gamma^*, C\rangle_F \tag{A.2}$$

where the parameters λ_1 and λ_2 control the marginal relaxation of the transport plan Γ^*, row-wise (source) and column-wise (target), respectively; and the parameter ϵ controls the sparsity of Γ.

In contrast to the balanced OT formulation (Eq. 2.2), this optimization requires three additional parameters. Higher λ_1 (λ_2) would impose a mass loss in the source (target) distributions. The ϵ is part of the sinkhorn optimization, which controls the sparsity of Γ, i.e. low values of ϵ result in a sparse transport plan Γ^*. For implementation details, we refer to POT documentation (https://pythonot.github.io/gen_modules/ot.unbalanced.html#ot.unbalanced.sinkhorn_unbalanced2).

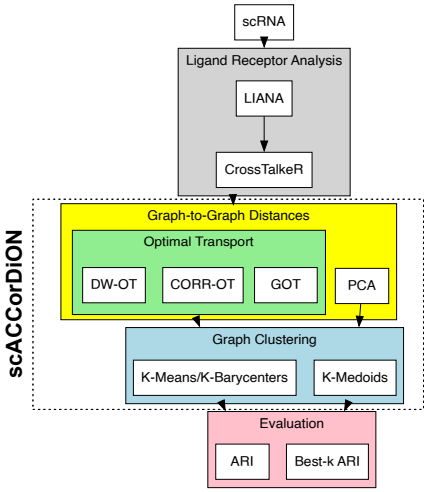


Figure A.1: Flowchart with the experimental design of this study and an overview of scACCorDiON: From the input scRNA cohort to the clustering benchmark

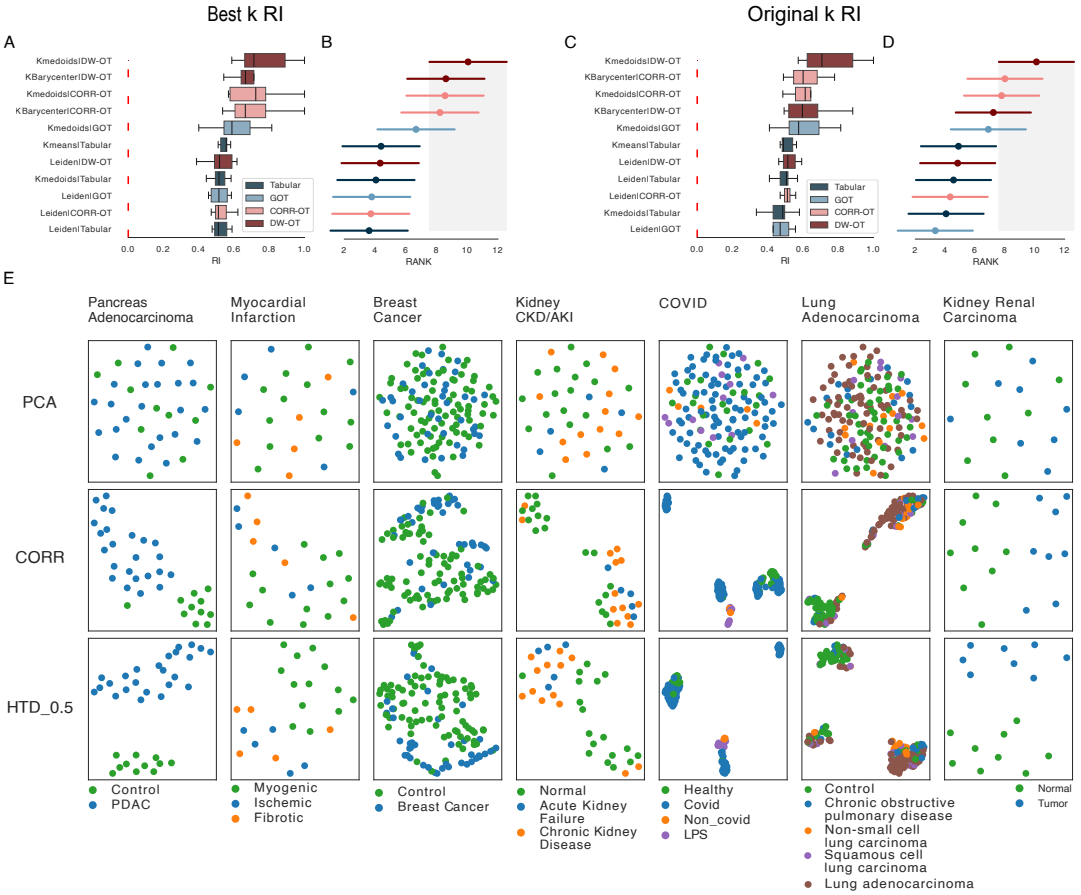


Figure A.2: Clustering Benchmark: A-Boxplots indicate the maximum RI value distribution (x-axis) for all evaluated methods over five scRNA-seq data sets. B-Ranking values (mean and std) for each method and data set regarding the maximum RI value. The highest ranking indicates the highest RI. The gray area indicates the 95% confidence interval of the Friedman and Nemenyi posthoc-test. Methods whose average values are not within the gray area have significantly lower rankings than the top-ranked methods. For both A and B, methods are ranked by average in decreasing order. C and D are the same as A and B for the RI estimates, with the number of clusters equal to the number of original labels. E-UMAP two-dimensional embeddings of the distance matrices estimated by `Tabular`, `CORR-OT`, and `DW-OT` for all evaluated data sets. Colors correspond to the original labels.

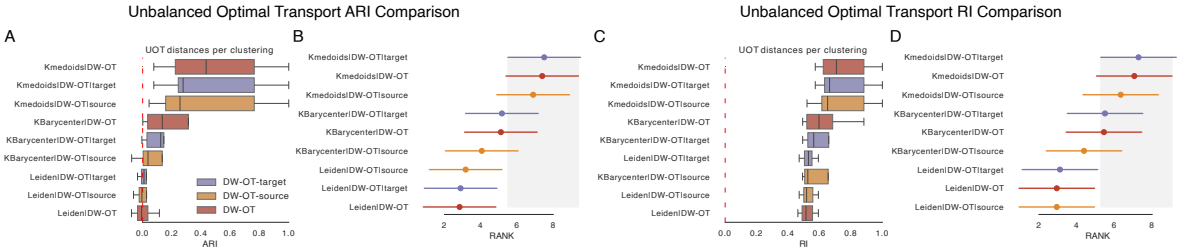


Figure A.3: A–Boxplots indicate the maximum RI value distribution (x-axis) for all evaluated methods over five scRNA-seq data sets. B–Ranking values (mean and std) for each method and data set regarding the maximum ARI value. The highest ranking indicates the highest ARI. The gray area indicates the 95% confidence interval of the Friedman and Nemenyi posthoc-test. Methods whose average values are not within the gray area have significantly lower rankings than the top-ranked methods. For both A and B, methods are ranked by average in decreasing order. C and D is the same as A and B for the RI estimates, with the number of clusters equal to the number of labels. Here, we evaluated two versions of the unbalanced algorithm with an increase in mass in either the source ($\lambda_1 = 1, \lambda_2 = 50$) or target distribution ($\lambda_1 = 50, \lambda_2 = 1$). We observe that for k-medoids and k-barycenter, the balanced OT presents equivalent performance of all other approaches.

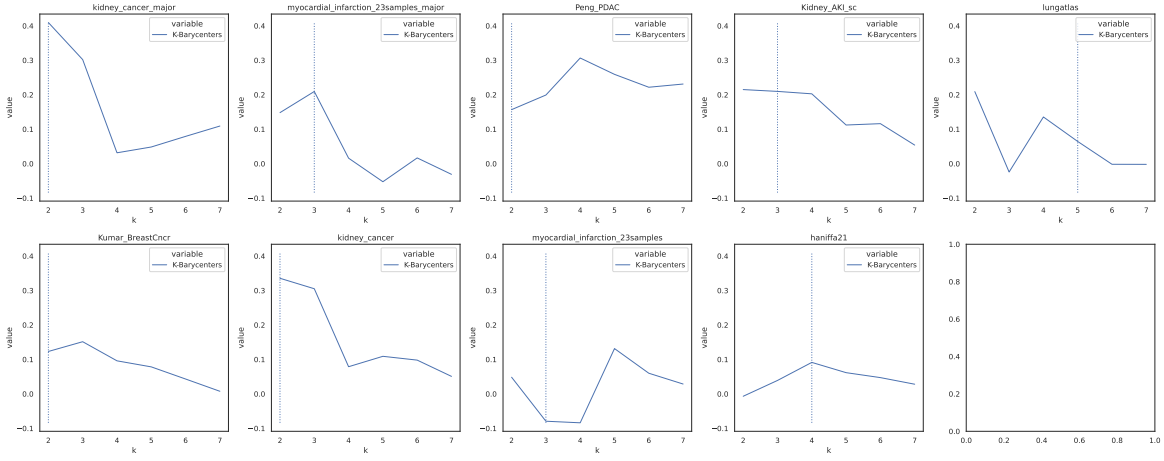


Figure A.4: Silhouette of different clustering methods using the DW-OT. Traced lines indicate the number of labels associated with each data set.

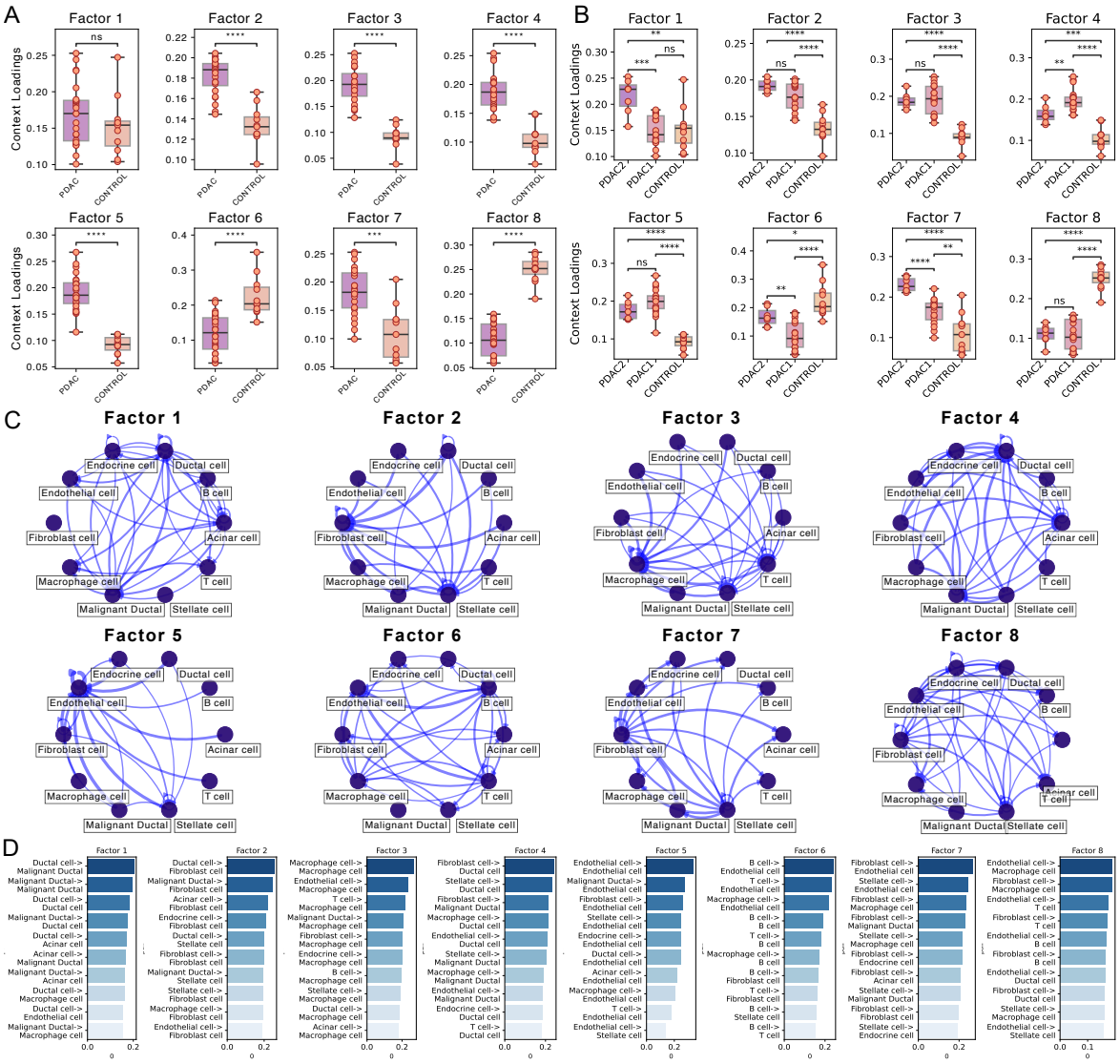


Figure A.5: Downstream analysis using Tensor-cell2cell: A-Statistical Comparison between the Context loading using the original disease labels; B-Statistical Comparison between the Context loading using the labels retrieved from scACCorDiON; C-Cell-Cell Communication Networks for each Factor. D-Top 10 highest cell-cell pairs' context loadings per factor.

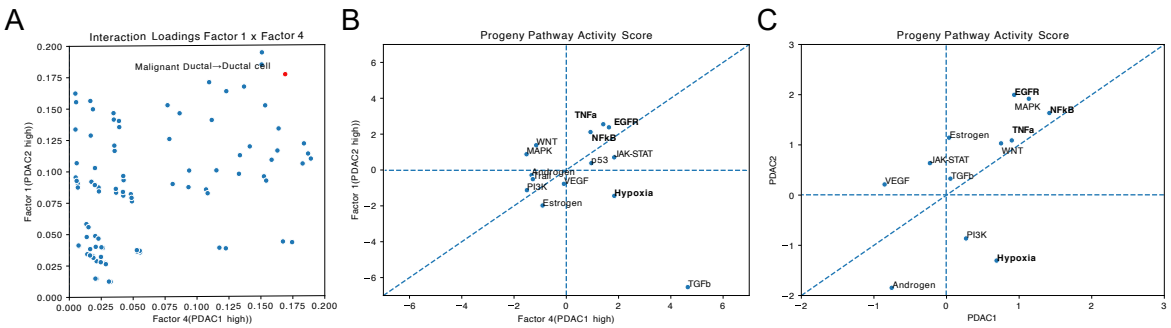


Figure A.6: Functional Analysis: A-Scatter plot with cell->cell interactions associated to PDAC2 related factor 1 (y-axis) and PDAC1 related factor 4 (x-axis). Both factors score the interaction Malignant Ductal cell -> Ductal Cells with high loading, which indicates its importance. B-Scatter plot with the progeny pathway activity for factor 1 (y-axis) vs factor 4 (x-axis). C-Scatter plot with progeny pathway activity for CrossTalkeR predicted PDAC 2 vs. PDAC 1 LR pairs for Malignant Ductal cell -> Ductal Cells.**Bold** indicates pathway activity found by both TensorCell2Cell and CrossTalkeR.