University of Windsor

Scholarship at UWindsor

Electronic Theses and Dissertations

Theses, Dissertations, and Major Papers

4-3-2024

Prediction of Cell-cell Communication from Spatial Transcriptomics Data Using a Long Short-term Memory Graph Neural Network

Karan Kashyap
University of Windsor

Recommended Citation

Kashyap, Karan, "Prediction of Cell-cell Communication from Spatial Transcriptomics Data Using a Long Short-term Memory Graph Neural Network" (2024). *Electronic Theses and Dissertations*. 9452.
https://scholar.uwindsor.ca/etd/9452

Prediction of Cell-cell Communication from Spatial Transcriptomics Data Using a Long Short-term Memory Graph Neural Network

By

Karan Kashyap

A Thesis
Submitted to the Faculty of Graduate Studies
through the School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science
at the University of Windsor

Windsor, Ontario, Canada

2024

Prediction of Cell-cell Communication from Spatial Transcriptomics Data Using a Long Short-term Memory Graph Neural Network

by

Karan Kashyap

APPROVED BY:

H. Wu
Department of Electrical and Computer Engineering

K. Selvarajah
School of Computer Science

L. Rueda, Advisor
School of Computer Science

March 21, 2024

DECLARATION OF CO-AUTHORSHIP / PREVIOUS PUBLICATION

I. Co-Authorship

I hereby declare that this thesis incorporates material that is result of joint research, as follows:

Chapter 2 of the thesis was co-authored with Akram Vasighizaker and Professor Luis Rueda. All authors collectively finalized the idea and engaged in follow-up discussions. Akram Vasighizaker refined the research idea and helped to drafting and reviewing the paper. The data collection, implementation, and writing and refining of the final draft were done by myself. Professor Luis Rueda supervised the whole project and provided initial thoughts on the main ideas, as well as reviewing and proofreading the final draft of the paper.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from each of the co-author(s) to include the above material(s) in my thesis.

I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

II. Previous Publication

This thesis includes original papers that have been previously submitted to conference, as follows:

Thesis chapter	**Publication title**	**Publication Status**
Chapter 2	K. Kashyap, A. Vasighizaker and L. Rueda, Prediction of Cell-cell Communication from Spatial Transcriptomics Data Using a Long Short-term Memory Graph Neural Network	Submitted

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as a graduate student at the University of Windsor.

III. General

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

Several studies are available that use gene expression data to infer cell-cell interactions. Nevertheless, most of these studies target intra-cellular interactions. The advent of spatial expression data paves the way for methodologies capable of deducing interactions, spanning both intra- and inter-cellular domains. However, spatial data also presents new challenges, including noisy and high-dimensional data and sparse representation. We propose a new model based on a graph neural network to predict cell-cell interaction from spatial data. Specifically, the study constructs a graph from the spatial data, forming the foundation for the model that combines the ability of Long Short-Term Memory and Graph Neural Network. The model's unique ability capitalizes on Long Short-Term Memory's sequence learning and Graph Neural Network's graph-based potential, designed to predict links within the spatial context. The model exhibits enhanced predictive capabilities through rigorous testing compared to similar approaches. Our investigation demonstrates that integrating our pipeline with the backward search technique yields the highest area under the curve (Area under Receiver Operating Characteristic curve) score. Furthermore, we have conducted a comparative analysis, juxtaposing this performance against two alternative approaches, SEAL (learning from Subgraphs, Embeddings and Attributes for Link prediction) and GCNG (Graph Convolutional Neural Networks for Genes). Our results demonstrate that integrating our pipeline with the backward search technique yields the highest Area under the ROC Curve score. The effectiveness of our approach is validated on two well-known datasets, seqFISH+ and Merfish, which capture the spatial intricacies of cellular communication.

Keywords: cell-cell interaction prediction, graph convolutional neural network, spatial transcriptomics, feature selection, LSTM.

DEDICATION

I would like to dedicate this thesis to my family.

ACKNOWLEDGEMENTS

This thesis has been made possible because of the constant support and resources made available by the University of Windsor.

I want to offer special thanks to my thesis supervisor, Dr. Luis Rueda, for his constant support and belief in me. He provided all the essentials that were demanded and required for conducting the experiments. Apart from that, he wonderfully mentored me at times regarding the research and kept me motivated.

I would also like to thank my external reader Dr. Huapeng Wu and internal reader Dr. Kalyani Selvarajah for being a part of my thesis committee and for their insightful comments and suggestions.

I am grateful for my parents' unwavering support and guidance throughout my journey. Also, I would like to express my gratitude to Akram Vasighizaker for her constant support and assistance throughout my research. Finally, I would like to give a special thanks to my friend Gautam Chandna who guided me through the initial phase of my degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
GNN	Graph Neural Network
LSTM	Long Short-Term Memory
AUROC	Area Under Receiver Operating Characteristic Curve
SEAL	Learning from Subgraphs, Embeddings, and Attributes for Link prediction
GCNG	Graph Convolutional Neural network approach for Genes
TP	True Positive
FN	False Negative
FP	False Positive
TN	True Negative
TPR	True Positive Rate
FPR	False Positive Rate

CHAPTER 1

Introduction

1.1 Basics Of Molecular Biology

Molecular biology is deals with the biological processes at the molecular level, understanding the interactions between and within various cellular components. In cell-cell interaction, molecular biology plays a pivotal role in unraveling the intricate mechanisms governing cell communication [1].

At the heart of cell-cell interactions are signaling molecules, often proteins or peptides, that transmit information between cells. These signaling events coordinate physiological processes, including development, immune response, and tissue homeostasis. Understanding these interactions at the molecular level provides insights into the regulatory networks that govern multicellular organisms [1]. One fundamental aspect is the concept of ligand-receptor interactions. Ligands are signaling molecules released by one cell, and receptors are proteins on the surface of another cell that recognize and bind to these ligands. This binding triggers a torrent of molecular events within the recipient cell, initiating a response. These interactions are highly specific, with each ligand binding to a corresponding receptor akin to a lock and key mechanism [17].

Molecular biology techniques, such as RNA sequencing, come into play to profile the gene expression patterns in interacting cells. RNA sequencing allows scientists to analyze the entire transcriptome, providing a comprehensive view of which genes are activated or repressed during cell-cell interactions. This information is critical for identifying the key players in signaling pathways and understanding how these interactions influence cellular behavior [1]. Moreover, advancements in molecular biology have led to the development of sophisticated tools like CRISPR-Cas9, enabling researchers to manipulate specific genes and observe the effects on cell behavior [11]. This precision in genetic editing is instrumental in deciphering the functional roles of various molecular components involved in cell-cell communication. Molecular biology provides the toolkit for investigating the molecular intricacies of cell-cell interactions. It sheds light on how cells communicate, coordinate their activities, and maintain the delicate balance required for the proper functioning of biological systems. Applying molecular biology techniques in studying cell-cell interactions is foundational for advancements in developmental biology, immunology, and cancer research [17].

1.1.1 Cells

Cells are the fundamental units of life, representing the basic structural and functional entities of all living organisms. In molecular biology, cells are the building blocks that form tissues, organs, and entire organisms. Understanding the intricacies of cell biology is essential for unraveling the complexities of molecular interactions and the functioning of living systems. At a basic level, cells are enclosed by a lipid bilayer comprehended as the cellular membrane, which separates the internal cellular environment from the external surroundings. This membrane serves as a selective barrier, regulating the passage of molecules in and out of the cell. Within the cell, various organelles, such as the nucleus, mitochondria, and endoplasmic reticulum, carry out specialized functions critical for the cell's survival and function. In multicellular organisms, cells communicate with one another via complex signaling networks. Cell-cell interactions are essential for coordinating processes such as development, immune response, and maintaining tissue homeostasis. These interactions involve exchanging signaling molecules, such as hormones, growth factors, and cytokines, between neighboring cells [1].

Distinct types of cells exist in the body, each specialized for particular functions. Stem cells, for example, can distinguish into different cell types, contributing to tissue repair and rejuvenation. Neurons transmit electrical signals in the nervous system, while immune cells are crucial in defending the body against pathogens. Refinements in molecular biology approaches have entitled scientists to explore different cell types and their functions. Single-cell RNA sequencing, for instance, enables gene expression profiling in individual cells, uncovering heterogeneity within cell populations. This heterogeneity is crucial for understanding the dynamic nature of cell populations and their responses to environmental cues [11].

Cells are the basic units of life, and their interactions form the foundation of biological processes. Molecular biology provides tools to study cells at the molecular level, revealing the intricacies of cellular functions, signaling pathways, and the dynamic nature of cell populations within multicellular organisms. The study of cells is pivotal for advancing our understanding of health, disease, and the fundamental principles of life [1].

1.1.2 Deoxyribonucleic acid and Ribonucleic acid

Deoxyribonucleic acid (DNA) and Ribonucleic acid (RNA) are fundamental molecules that play pivotal roles in storing and transmitting genetic information, serving as the building blocks of life. DNA, often described as the ”genetic blueprint,” contains the instructions necessary for all living organisms' growth, development, and functioning. DNA possesses a unique double-helix network of two long strands that coil about each other in a right-handed spiral. The structure is stabilized by hydrogen bonds between complementary pairs of nucleobases. The four types of nucleobases found in DNA are adenine (A), cytosine (C), guanine (G), and thymine (T). Adenine couples with thymine, and cytosine couples with guanine, forming the basis for the complementary nature of DNA strands [1].

The information encoded in DNA is transferred into RNA through a process known as transcription. RNA, like DNA, is composed of nucleotides but is typically single-stranded. The three main types of RNA are messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). mRNA serves as a temporary copy of the genetic code, carrying it from the DNA in the cell nucleus to the cellular machinery, where proteins are synthesized. The nucleotide structure of RNA is similar to that of DNA, with a phosphate group, ribose sugar, and nucleobases. However, in RNA, thymine is substituted by uracil (U). RNA polymerase reads the DNA template during transcription and synthesizes a complementary RNA strand, incorporating uracil instead of thymine. The functional significance of DNA lies in its role as a hereditary molecule, passing genetic information from one generation to the next. It undergoes processes like replication, ensuring faithful transmission of genetic material during cell division. Mutations in DNA can lead to genetic variations, providing the raw material for evolution [1].

Conversely, RNA plays a crucial role in the synthesis of proteins, acting as an intermediary between the genetic code in DNA and the actual protein production. The information carried by mRNA is translated by ribosomes, with the help of tRNA, into a specific sequence of amino acids, forming proteins essential for cellular structure and function. DNA and RNA are intricately connected molecules, with DNA serving as the stable repository of genetic information and RNA facilitating the dynamic processes of gene expression and protein synthesis. Together, they orchestrate the molecular ballet that defines life's complexity and diversity [9].

1.1.3 Ligands

In molecular biology, ligands bind to a specific site on a target molecule, often a larger biomolecule, to form a complex. Ligand-receptor interactions are fundamental in cellular communication and are pivotal in various physiological processes. Understanding the nature of ligands and their interactions is crucial for unraveling the intricacies of signaling pathways and molecular events within cells.

Ligands can vary widely in structure and function. They can be ions, small molecules, or large proteins. Hormones, neurotransmitters, and growth factors are typical examples of ligands involved in cell signaling. The specificity of ligand binding is often determined by the shape and chemical properties of both the ligand and the binding site on the target molecule.

Ligands are not limited to interactions with cell surface receptors. Some ligands are designed to act within the cell, binding to receptors located inside the cell. In this case, ligands often must pass through the cell membrane to reach their target receptors. Understanding ligand-receptor interactions is crucial in drug development. Many pharmaceutical drugs work by acting as ligands, either mimicking the action of endogenous ligands or blocking specific interactions. Designing drugs that selectively target certain receptors allows for developing therapies with fewer side effects [17].

Ligands are critical players in cellular communication, mediating essential processes in health and disease. The study of ligand-receptor interactions provides insights into the molecular mechanisms governing cellular responses and opens avenues for therapeutic interventions in various fields, including medicine and biotechnology [6].

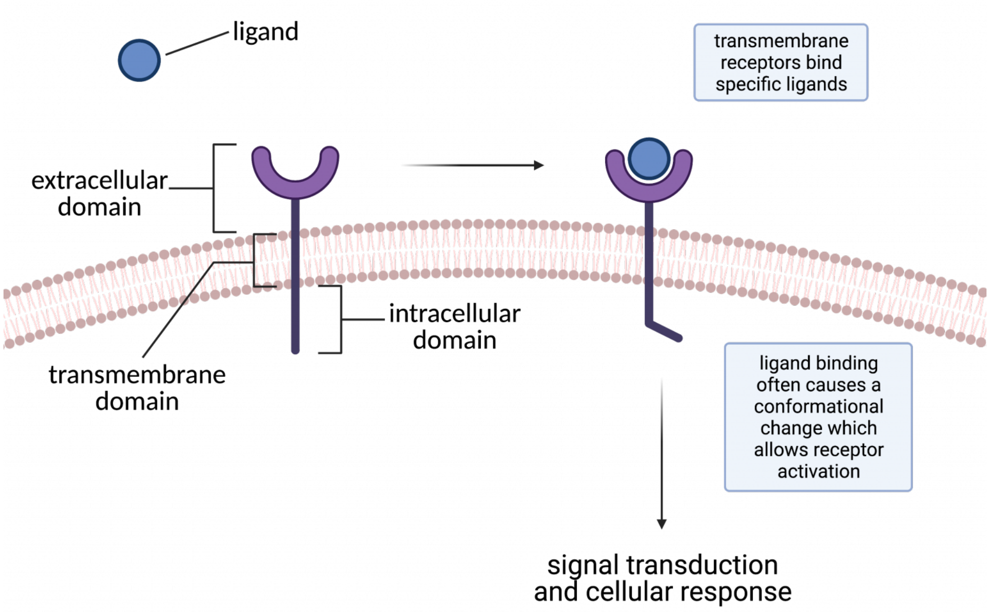

ligand
transmembrane receptors bind specific ligands
extracellular domain
intracellular domain
transmembrane domain
ligand binding often causes a conformational change which allows receptor activation
signal transduction and cellular response

Fig. 1.1.1: The basic functionality of ligand and receptor for cell-cell communication [3].

1.1.4 Receptors

The most well-known class of ligands is signaling molecules that attach to receptors on the surface of cells. These interactions initiate signaling cascades that regulate various cellular responses. For instance, neurotransmitters binding to receptors on neurons trigger nerve impulses, while hormones binding to cell surface receptors can influence gene expression and cell behavior. Receptors are molecules, often proteins, that receive signals from the external environment or other cells and initiate a specific cellular response. Receptors play a central role in cell communication, allowing cells to respond to their surroundings and coordinate various physiological processes [6].

When a ligand binds to a receptor, it triggers a series of events known as signal transduction, leading to a cellular response. This process often involves the activation of intracellular proteins, changes in gene expression, or alterations in cell behavior. The specificity of receptor-ligand interactions is critical for the proper functioning of cells. Receptors exhibit high specificity, recognizing and binding to particular ligands with precise affinity. This specificity ensures that cells respond appropriately to specific signals [17].

Receptors are involved in diverse cellular processes, including:

- Cell Growth and Differentiation: Receptors regulate cell division and differentiation in response to growth factors.
- Immune Response: Immune cells use receptors to detect signals from pathogens and other immune cells.
- Neuronal Signaling: Neurotransmitter receptors are crucial in transmitting signals between nerve cells [6].

1.2 RNA Sequencing

RNA sequencing (RNA-Seq) is a powerful molecular biology technique that provides a comprehensive and quantitative transcriptome analysis, encompassing all the RNA molecules within a cell or tissue at a specific moment. It has revolutionized the study of gene expression by enabling researchers to profile RNA molecules, measure their abundance, identify novel transcripts, and understand the dynamic nature of gene regulation. The RNA-Seq process begins with isolating RNA from the biological sample of interest, such as cells or tissues. This RNA can be a mix of various types, including messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and non-coding RNA. The next step involves converting the isolated RNA into a complementary DNA (cDNA) library through reverse transcription. This cDNA library represents a snapshot of the RNA present in the sample [4].

Once the cDNA library is prepared, high-throughput sequencing technologies sequence the cDNA fragments. The most common sequencing platforms include Illumina, Ion Torrent, and PacBio. These platforms generate millions or even billions of short DNA sequences, or ”reads,” in a massively parallel fashion. Each read corresponds to a fragment of cDNA derived from an RNA molecule in the original sample. After sequencing, the resulting data must be processed and analyzed. Depending on the experimental design, Bioinformatics tools align the short reads to a reference genome or transcriptome. This alignment step allows researchers to map the sequenced fragments back to their genomic or transcriptomic origin, providing insight into the expression levels of specific genes and transcripts [4]. Quantification of gene expression is a primary goal of RNA-Seq analysis. The number of reads mapped to each gene or transcript serves as a measure of its abundance. This information enables researchers to compare gene expression levels across different conditions, such as healthy and diseased tissues, before and after a specific treatment. Differential gene expression analysis identifies up or downregulated genes under specific experimental conditions.

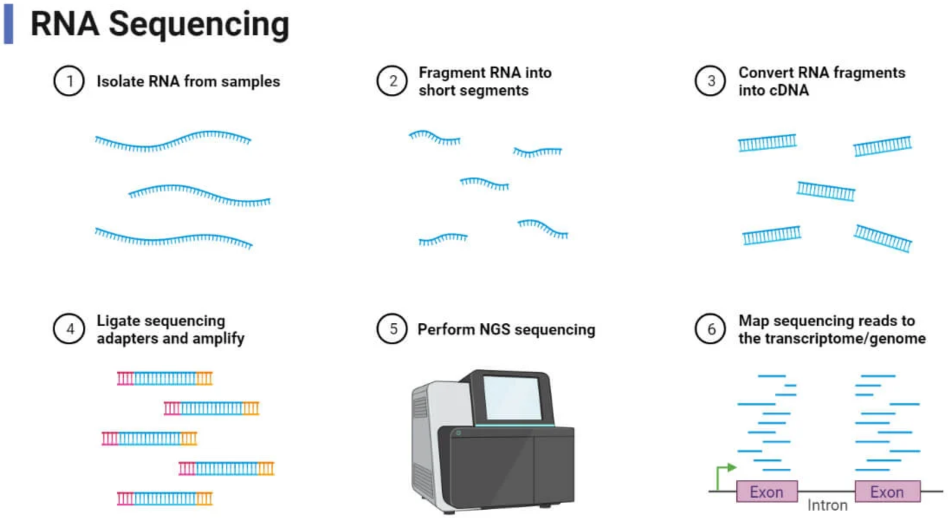

RNA Sequencing
1 Isolate RNA from samples
2 Fragment RNA into short segments
3 Convert RNA fragments into cDNA
4 Ligate sequencing adapters and amplify
5 Perform NGS sequencing
6 Map sequencing reads to the transcriptome/genome
Exon
Intron
Exon

Fig. 1.2.1: the basic steps involved in generation of RNA-sequencing dataset [4].

RNA-sequencing is not limited to measuring gene expression alone; it can also capture alternative splicing events, identify novel transcripts, and detect non-coding RNAs. Alternative splicing, a process where a single gene can produce multiple mRNA isoforms, contributes to the diversity of proteins generated from a limited number of genes. RNA-Seq permits researchers to specify and quantify additional splice variants, providing a more nuanced understanding of gene function. The ability of RNA-sequencing to detect novel transcripts and non-coding RNAs has expanded our understanding of the complexity of the transcriptome. It has revealed the presence of long non-coding RNAs and small non-coding RNAs that play crucial roles in regulating gene expression, cellular processes, and disease. RNA sequencing is a transformative technology that has revolutionized the field of genomics. Its ability to provide a detailed and quantitative snapshot of the transcriptome allows researchers to explore gene expression, alternative splicing, and non-coding RNA biology with unprecedented depth. The information derived from RNA-sequencing experiments contributes significantly to our understanding of cellular processes, development, and the molecular mechanisms underlying various diseases [9].

1.3 Spatial Transcriptomics Data

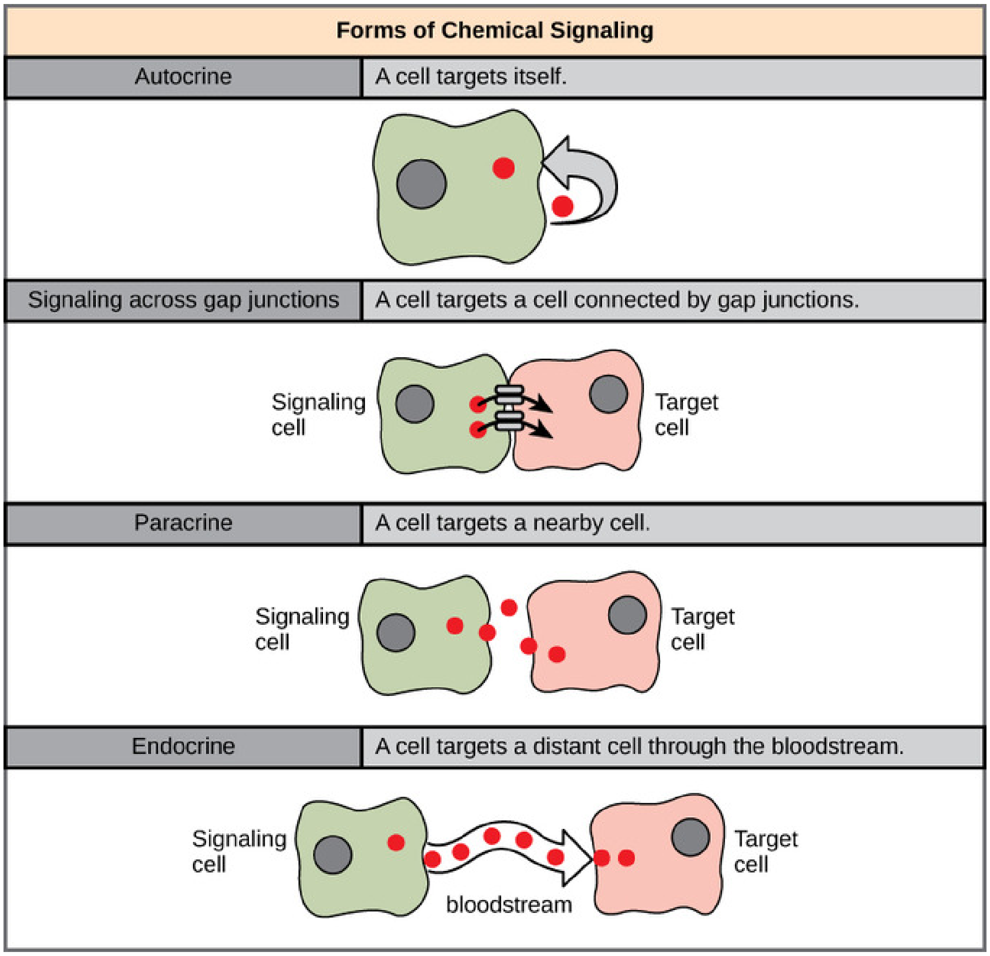
Spatial transcriptomics is a cutting-edge field in genomics that goes beyond traditional gene expression analysis by preserving the spatial context of gene activity within tissues. Unlike conventional transcriptomics approaches, which provide insights into the abundance of mRNA molecules in a bulk tissue sample, spatial transcriptomics seeks to unravel the intricate patterns of gene expression about the spatial organization of cells. This spatial dimension is critical for understanding the complex architecture of tissues and the nuanced interplay between different cell types [5].

In spatial transcriptomics, the data generated includes information about the identity and abundance of transcripts and retains the positional information of where these transcripts are located within the tissue. This is achieved through innovative technologies that allow researchers to map the transcriptome onto the spatial coordinates of the tissue. Techniques like seqFISH (Sequential Fluorescence In Situ Hybridization) and MERFISH (Multiplexed Error-Robust Fluorescence In Situ Hybridization) enable the visualization of gene expression at the single-cell level while preserving the spatial relationships between neighboring cells [15].

The spatial transcriptomics workflow typically involves the collection of tissue samples, followed by the preservation of spatial information during the extraction of RNA. High-throughput sequencing techniques, such as RNA-Seq, are then employed to capture the transcriptome of individual cells. Importantly, spatial information is retained by preserving the spatial coordinates of the cells during the library preparation and sequencing steps. One of the critical advantages of spatial transcriptomics is its ability to provide a holistic view of the gene expression landscape within tissues. This approach allows researchers to identify the types of cells present and their spatial distribution and interactions [5]. For instance, spatial transcriptomics can unveil the communication networks between different cell types in cell-cell interactions, shedding light on signaling cascades and molecular dialogues within the tissue microenvironment. The spatially resolved gene expression data obtained from spatial transcriptomics experiments has diverse applications. It can aid in identifying spatially restricted cell populations, elucidate the role of specific genes in defined tissue regions, and provide insights into how cells coordinate their activities in a spatially organized manner. This spatially aware transcriptomic data is precious in areas such as developmental biology, neuroscience, and research on cancer, where the spatial organization of cells is crucial for understanding physiological processes and disease mechanisms [5].

1.4 Cell-cell communication

Cell-cell communication is a fundamental biological process that allows cells to harmonize their activities, respond to environmental cues, and participate in the intricate orchestration of physiological processes. This communication is vital for the correct functioning and homeostasis of multicellular organisms, where diverse cell types function concurrently in a coordinated way to support the overall health and functionality of tissues and organs. At the core of cell-cell communication is the exchange of signals between neighboring cells. These signals can take various forms, including chemical, electrical, and mechanical signals, each playing a specific role in different biological contexts [7]. The communication between cells is not a passive process; instead, it involves a dynamic interplay where cells send and receive signals, allowing them to sense their environment and respond accordingly. Chemical signaling is a predominant mode of cell-cell communication. It involves the release of signaling molecules, often called ligands, by one cell and the recognition of these molecules by receptors on the surface of a neighboring cell. This interaction triggers a cascade of events within the recipient cell, leading to a specific response. Ligands can be small molecules, proteins, or even nucleic acids, and they bind to receptors with high specificity, ensuring the precision of the cellular response [6].



Forms of Chemical Signaling

Autocrine	A cell targets itself.
Signaling across gap junctions	A cell targets a cell connected by gap junctions.
Paracrine	A cell targets a nearby cell.
Endocrine	A cell targets a distant cell through the bloodstream.

Fig. 1.4.1: Different ways in which the cells can communicate [3]

There are different ways in which the cell can communicate as illustrated in Figure 1.4.1.

Autocrine Signaling: Autocrine signaling is a cell communication mode where ligands a cell produces to bind to receptors on the same cell that secreted them. In this self-stimulatory process, the cell signals itself, often to regulate its activity or maintain a certain cellular state. This form of signaling allows cells to respond to their secretions, enabling a fine-tuned control mechanism. For instance, a cell might release a signaling molecule, such as a growth factor, and then detect and respond to that signal via its receptors. Autocrine signaling is vital for various cellular functions, including growth, differentiation, and immune responses.

Paracrine Signalling: In paracrine signaling, cells release signaling molecules (ligands) that affect nearby target cells in the immediate environment. Ligands travel short distances to interact with receptors on neighboring cells, influencing their behavior or activities. This type of communication is crucial for coordinating cellular activities within specific tissues or microenvironments. For example, immune cells release signaling molecules to alert nearby cells of an infection, triggering a localized response. Paracrine signaling regulates tissue repair, inflammation, and neuronal communication.

Signaling Across Gap Junctions: Gap junctions provide a direct physical connection between adjacent cells, allowing the passage of ions, small molecules, and signaling molecules directly from one cell to another. This direct cell-to-cell communication mechanism is known as signaling across gap junctions. It enables rapid and coordinated responses among connected cells. This form of communication is common in tissues that require synchronized activity, such as cardiac muscle; through gap junctions, cells share information, collectively coordinating their functions.

Endocrine Signalling: Endocrine signaling involves the release of signaling molecules, typically hormones, into the bloodstream by endocrine glands. These molecules travel through the circulatory system to reach distant target cells in various body parts. Unlike autocrine, paracrine, and gap junction signaling, endocrine signaling has no distance limitations. It enables cells to communicate across long distances, allowing for systemic coordination and regulation of physiological processes. For instance, the thyroid gland secretes hormones into the bloodstream that affect target cells throughout the body, influencing metabolism and energy balance [3].

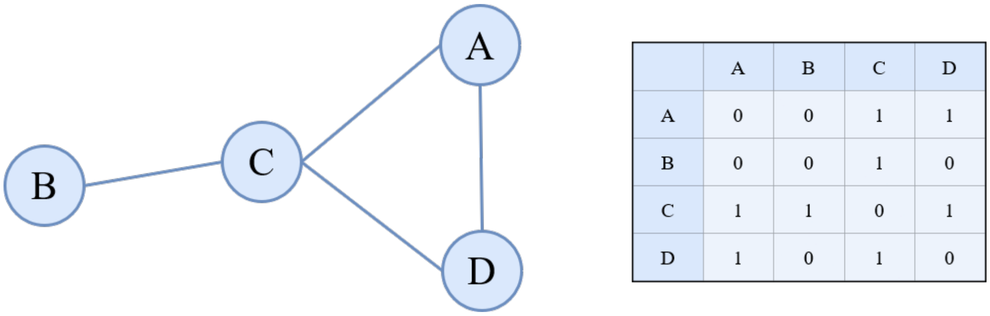
Cell-cell communication is closely intertwined with cell-cell interaction, where the physical and molecular contacts between cells influence their behavior and function. Cell-cell interactions encompass a broad spectrum of phenomena, ranging from direct physical connections, such as gap junctions that permit direct communication between the cytoplasm of neighboring cells, to indirect signaling through diffusible molecules like growth factors and hormones. In the context of tissues and organs, cell-cell interactions are vital for processes like development, immune response, and tissue repair [2]. For example, during embryonic development, precise cell-cell communication guides the formation of tissues and organs, ensuring that cells differentiate into the right cell types and migrate to their designated locations. Cell-cell communication coordinates the response to pathogens in the immune system, allowing different immune cells to work together to mount an effective defense. Cell-cell communication is also integral to pathological processes, including cancer. Dysregulation of communication pathways can lead to uncontrolled cell growth, invasion of surrounding tissues, and metastasis. Understanding the nuances of cell-cell communication in both healthy and diseased states is crucial for developing targeted therapies that modulate these interactions to restore normal cellular function [6].

1.5 Graphs

Graphs comprise nodes (vertices) (V) and edges (E). Nodes symbolize entities, and edges depict connections or relationships between them. In the context of cell-cell interaction studies, nodes represent individual cells ($c_i \in V$), while edges represent pairwise interactions ($e_{ij} \in E$) between cells. This graph-based representation, denoted by $G = (V, E)$, effectively captures the inherent relational structure within complex systems, offering a versatile framework for modeling and analyzing diverse interactions, including those within molecular signaling pathways. The edges act as conduits of information, conveying the connections between entities and providing a visual and computational means to dissect and interpret the intricate relationships within biological networks.[14].

1.5.1 Adjacency Matrix

The adjacency matrix, given by $A = \{a_{ij}\}$, is a square matrix that encodes the connections within a graph. In the context of cell-cell interaction graphs, every row and column of the adjacency matrix represents a cell, and the element a_{ij} at row i and column j indicates whether there is a direct interaction between cell i and cell j. A value of 1 signifies a connection (edge) between the corresponding cells, while 0 indicates the absence of a direct interaction. This binary matrix provides a concise and computationally efficient way to capture the network's structure, making it a fundamental tool in analyzing cell communication patterns within spatial transcriptomics datasets.[7].



	A	B	C	D
A	0	0	1	1
B	0	0	1	0
C	1	1	0	1
D	1	0	1	0

Fig. 1.5.1: An exemplar adjacency matrix that illustrates the interactions between nodes in a network derived from the data. 1 denotes a relationship, 0 denotes none.

1.5.2 Directed and Undirected Graphs

Directed and undirected graphs represent relationships between entities, denoted as $G = (V, E)$. In a directed graph, edges ($e_{ij} \in E$) have a defined direction, signifying a one-way relationship between nodes ($v_i, v_j \in V$). This analogy holds true for directed cell-cell interactions, where the influence or signaling flows from one cell to another. Conversely, undirected graphs denote symmetric relationships without a specified direction, reflecting mutual interactions. In cell-cell interaction graphs, undirected edges represent bidirectional signaling. These graph types are critical in spatial transcriptomics, where the directionality of molecular signaling and the symmetry of interactions play a significant role in understanding complex cellular communication networks. Additionally, the adjacency matrix $A = \{a_{ij}\}$ encodes these relationships, where $a_{ij} = 1$ indicates a connection among the nodes v_i and v_j, and $a_{ij} = 0$ indicates their absence. [8].

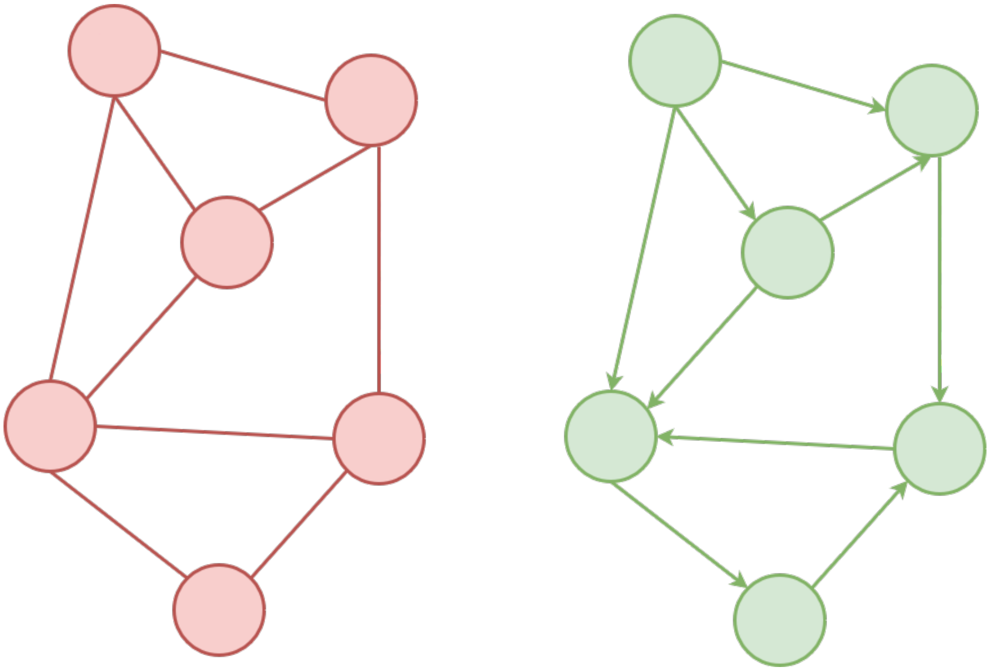


Fig. 1.5.2: Shows the difference between the directed and undirected graphs.

1.5.3 Attributed graphs

Attributed graphs incorporate additional attributes associated with nodes or edges, enriching the basic graph structure. In cell-cell interactions, attributed graphs can represent diverse features like gene expression levels, cell types, or signal strengths. Nodes and edges carry attribute vectors, offering a multi-dimensional perspective on cellular relationships. This is particularly valuable in spatial transcriptomics, where attributing nodes with spatial information or molecular attributes enhances the understanding of intricate cellular networks. Attribute-rich graphs provide a comprehensive framework to model complex biological phenomena, enabling nuanced analyses of cell interactions and spatial organization [8].

1.6 Machine Learning

Machine learning has recently played a key role in analyzing complex biological data [7] [14]. It can be applied to identify patterns in gene expression profiles, predict cell-cell communication networks, and uncover hidden relationships within high-dimensional spatial transcriptomics datasets. Machine learning models, including graph-based neural networks, excel at capturing intricate dependencies in biological data, providing valuable insights into cellular behavior and interaction dynamics [14]. The application of machine learning in biology and medicine is expanding rapidly, contributing to advancements in personalized medicine, drug discovery, and understanding of complex biological processes. As technologies for data generation in the life sciences continue to evolve, machine learning is poised to play an increasingly critical role in unraveling the complexities of cell-cell interactions and advancing our understanding of biological systems [2].

1.6.1 Graph Neural Network

Graph Neural Networks (GNNs) are machine learning models designed to handle data with a graph or network structure, making them particularly powerful for tasks involving relationships and interactions. In cell-cell interaction studies, where cellular relationships form a complex network, GNNs are instrumental. Unlike conventional neural networks that use grid-like data such as images or sequences, GNNs can effectively capture the intricate dependencies present in graph-structured data. These networks extend neural network architectures to process and analyze information from nodes (representing entities, such as cells) and edges (representing connections, such as interactions between cells) in a graph [18].

Graph Neural Networks (GNNs) represent a transformative paradigm in machine learning, specifically designed to handle data structured as graphs or networks ($G = (V, E)$). Their unique architecture, leveraging message-passing mechanisms, allows them to effectively capture and propagate information across nodes and edges, making them adept at addressing tasks involving intricate relationships and interactions.

GNNs have proven particularly influential in cell-cell interaction studies, where cellular relationships naturally form a complex network [18]. Unlike traditional neural networks excelling in handling grid-like data (e.g., images, sequences), GNNs are purpose-built to unravel the complexities embedded in **graph-structured data** like cell-cell interaction networks. This distinctive feature positions them as powerful tools for understanding and modeling relationships in various domains.

For instance, in cell-cell interaction analysis, GNNs can operate on a graph representation G where nodes ($v_i \in V$) represent individual cells and edges ($e_{ij} \in E$) signify interactions between them. The adjacency matrix $A = \{a_{ij}\}$ encodes these connections, where $a_{ij} = 1$ indicates a connection between nodes v_i and v_j, and $a_{ij} = 0$ indicates their absence. By exploiting the inherent structure of the graph and propagating information across nodes and edges, GNNs can effectively learn to predict cell-cell interactions based on node and edge features, GNNs can predict the likelihood of interaction between two cells, potentially uncovering novel or hidden interactions within the network. Additionally, it can classify cell types; GNNs can leverage node features and information from neighboring cells to classify cell types within the network, providing insights into cellular composition and organization. These are just a few examples of how GNNs revolutionize our understanding of cell-cell interactions and complex biological networks. Their ability to handle graph-structured data effectively makes them a powerful tool for various tasks like network analysis, drug discovery, disease classification, and biomarker discovery in computational biology and beyond [14].

In cell-cell interactions, where the relationships between individual cells weave a complex network, GNNs provide a versatile and efficient solution. The nodes in the graph typically represent entities, such as cells, while the edges signify connections or interactions between these entities. This graph-based representation allows GNNs to capture the nuanced dependencies and contextual intricacies inherent in cellular interactions. Traditional machine learning models need help to navigate the intricate web of relationships in such scenarios. However, GNNs excel in discerning patterns and extracting meaningful information from these graph structures [15].

The core strength of GNNs lies in their ability to aggregate information from nodes and edges within a graph. By leveraging this comprehensive understanding of the relationships between entities, GNNs can make informed predictions or classifications. This becomes particularly crucial in domains like biology, where individual cells' behavior and interactions play a pivotal role in overall system dynamics.

One notable aspect of GNNs is their capacity for representation learning. Through the iterative processing of node and edge information, GNNs can generate meaningful embeddings that capture the underlying structure and patterns in the graph. This representation learning enables GNNs to effectively encode complex relationships, making them invaluable in predicting cell behavior, understanding disease mechanisms, or optimizing biological processes.

The adaptability of GNNs extends beyond biological domains to various fields where relationships are best represented as graphs. Applications span social network analysis, recommendation systems, cybersecurity, transportation networks, and more. GNNs provide a versatile framework that accommodates diverse data structures, offering a powerful solution for understanding and leveraging complex relationships in real-world scenarios. Graph Neural Networks stand as a groundbreaking development in machine learning, specifically tailored for tasks involving graph-structured data. Their adeptness in capturing complex relationships, as exemplified in cell-cell interaction studies, showcases their transformative potential in understanding intricate systems. With applications spanning multiple domains, GNNs represent a pivotal advancement in pursuing more accurate and insightful machine learning models [18].

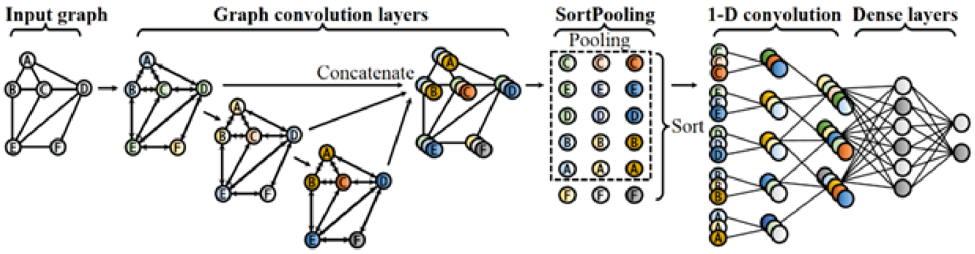


Fig. 1.6.1: The structure of Deep Graph Convolutional Neural Network (DGCNN), the underline GNN used in SEAL [18].

In the study of spatial transcriptomics, GNNs can be employed to model the spatial organization of cells, identify signaling pathways, and predict communication patterns within tissues. By leveraging the inherent structure of cellular networks, GNNs contribute significantly to unraveling the complexities of cell-cell interactions and enhancing our understanding of biological systems [15].

1.6.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM), denoted as $LSTM(x_t, h_{t-1}, c_{t-1})$, represents a recurrent neural network (RNN) architecture designed to overcome challenges in capturing long-term dependencies within sequential data. This proves crucial in studying cell-cell interactions, where understanding sequences is critical. Traditional RNNs suffer from limitations in handling long sequences, but LSTMs address this by introducing memory cells (c_t) and gating mechanisms:

- **Input gate** (i_t)**:** Regulates new information flow (x_t) into the cell.
- **Forget gate** (f_t)**:** Controls information to be discarded from the previous cell state (c_{t-1}).
- **Cell update:** Combines forgotten and new information using cell state updates.
- **Output gate** (o_t)**:** Controls the information output (h_t) based on the updated cell state.
- **Hidden state update:** Combines the updated cell state and output gate to produce the new hidden state.

These gating mechanisms enable LSTMs to selectively remember and utilize information across long timeframes, making them adept at modeling temporal sequences in cell-cell interaction studies, gene expression analysis, and other applications with sequential data. [12].

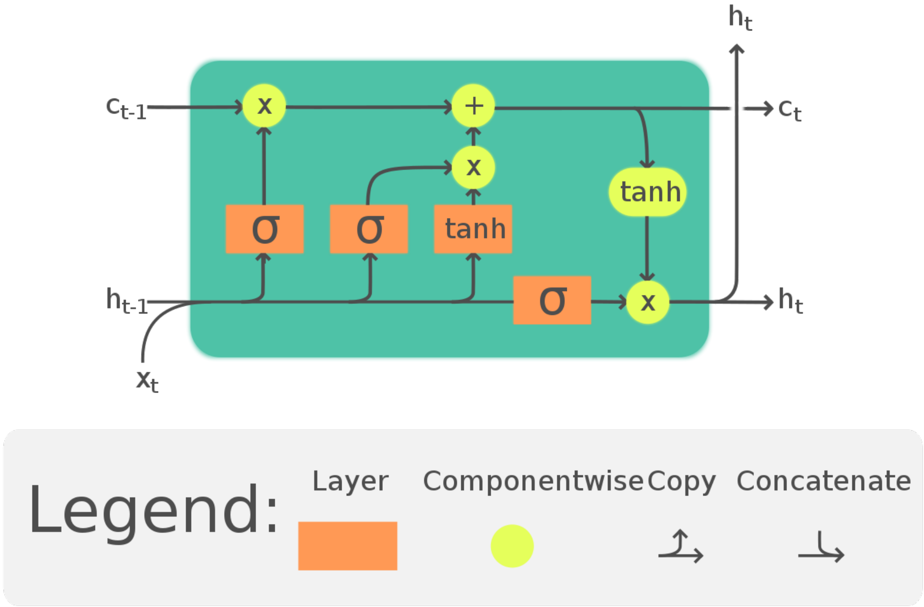


Fig. 1.6.2: The basic structure of LSTM [12].

As illustrated in 1.6.2, Long Short-Term Memory (LSTM) networks have significantly advanced recurrent neural networks (RNNs) designed to combat the challenges inherent in handling long-term dependencies – a limitation plaguing traditional RNNs. The architectural intricacies of LSTMs involve four interconnected layers: the input gate, forget gate, cell state, and output gate. These layers collaborate to manage the information flow and control the memory retention within the cell. The LSTM cell produces two primary outputs: the cell state, acting as the memory reservoir capable of carrying information across extended sequences, and the cell output, which is derived from the cell state and transmitted to the subsequent hidden layer.

Unlike conventional RNNs, LSTMs incorporate three logistic sigmoid gates – output, forget, and input gates – and a hyperbolic tangent (tanh) layer. The sigmoid gates play a pivotal role in determining the relevance of information and deciding which parts of the data should be preserved and which should be discarded. The tanh layer, on the other hand, processes selected input information, generating values in the range of -1 to 1. This combination of gates and layers enables LSTMs to address the challenges associated with learning long-term dependencies.

The basic functioning of LSTMs involves several key steps. First, the input gate processes the current input, deciding which information should be stored in the cell state. Simultaneously, the forget gate assesses the previous cell state, determining which information should be discarded. The cell state is then updated by combining the relevant information from both the input and forget gates. This mechanism allows LSTMs to selectively remember or forget information over time, enabling the network to maintain context over extended sequences. Finally, the output gate decides what information from the cell state should be passed to the next hidden layer, and both the cell output and the updated cell state are transmitted to the subsequent layer in the network.

The advantages of LSTMs over traditional RNNs are noteworthy. One of the primary benefits lies in their ability to handle long-term dependencies effectively, a critical improvement over the vanishing gradient problem observed in RNNs. The introduction of gated mechanisms in LSTMs provides finer control over the preservation and forgetting of contextual information. Additionally, LSTMs nearly eliminate the vanishing gradient problem, contributing to more stable and efficient training of deep networks. Furthermore, LSTMs exhibit flexibility in handling diverse data types, including noise, distributed representations, and continuous data. Unlike traditional models like hidden Markov models (HMMs), LSTMs do not require a fixed number of states from the outset, offering adaptability in various applications. Long Short-Term Memory Networks have arisen as a powerful solution for processing sequential data, mainly where long-term dependencies are crucial. Their intricate architecture, encompassing gated mechanisms and memory management, enables them to effectively capture and retain information over extended sequences, addressing critical limitations of traditional RNNs [12].

In studying cell-cell interactions, LSTM aids in modeling the sequential nature of gene expression patterns. This is essential for understanding how cells communicate and respond to their microenvironment over time. LSTMs learn patterns in sequential data, enabling them to discern meaningful signals in complex biological processes [12].

1.7 Performance Metrics

Performance metrics are crucial as they provide quantitative benchmarks to assess the effectiveness of predictive models. They offer insights into the model's accuracy and ability to discern between different classes, guiding researchers in selecting and refining models for optimal performance in cell-cell interaction predictions [10].

The prediction of cell-cell communication can help uncover new information in understanding biological processes and discover some new interactions among or within cells. This can be difficult due to complexity of the data that arises due to high dimensionality and sparse nature of the data. Keeping these points in mind, we have chosen two critical metrics for the evaluation of our model – Accuracy and AUC.

1.7.1 Accuracy

Accuracy is a rudimentary metric that measures the overall correctness of a model's predictions. In cell-cell interaction prediction, accuracy quantifies the proportion of correctly predicted interactions against the total predictions. It is a straightforward measure, clearly indicating how well the model is performing across the entire dataset. However, accuracy might be limited in scenarios with imbalanced datasets, where one class (interaction or non-interaction) dominates, potentially leading to skewed evaluations [10].

Mathematically, it can be expressed as:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

1.7.2 Area Under the Curve

Area under the curve (AUC) is a performance metric widely employed in binary classification problems, particularly useful when assessing the predictive capabilities of models in cell-cell interaction studies. The ROC (Receiver Operating Characteristic) curve, from which AUC is derived, plots the trade-off between true positive rates and false positive rates at various thresholds. AUC quantifies the model's ability to distinguish between positive and negative instances, regardless of the chosen threshold. In spatial transcriptomics, where accurate discrimination between cell-cell interactions and non-interactions is crucial, AUC provides a nuanced assessment, especially when class distribution is imbalanced. A high AUC indicates superior discriminatory power, showcasing the model's proficiency in distinguishing between different interaction classes [15].

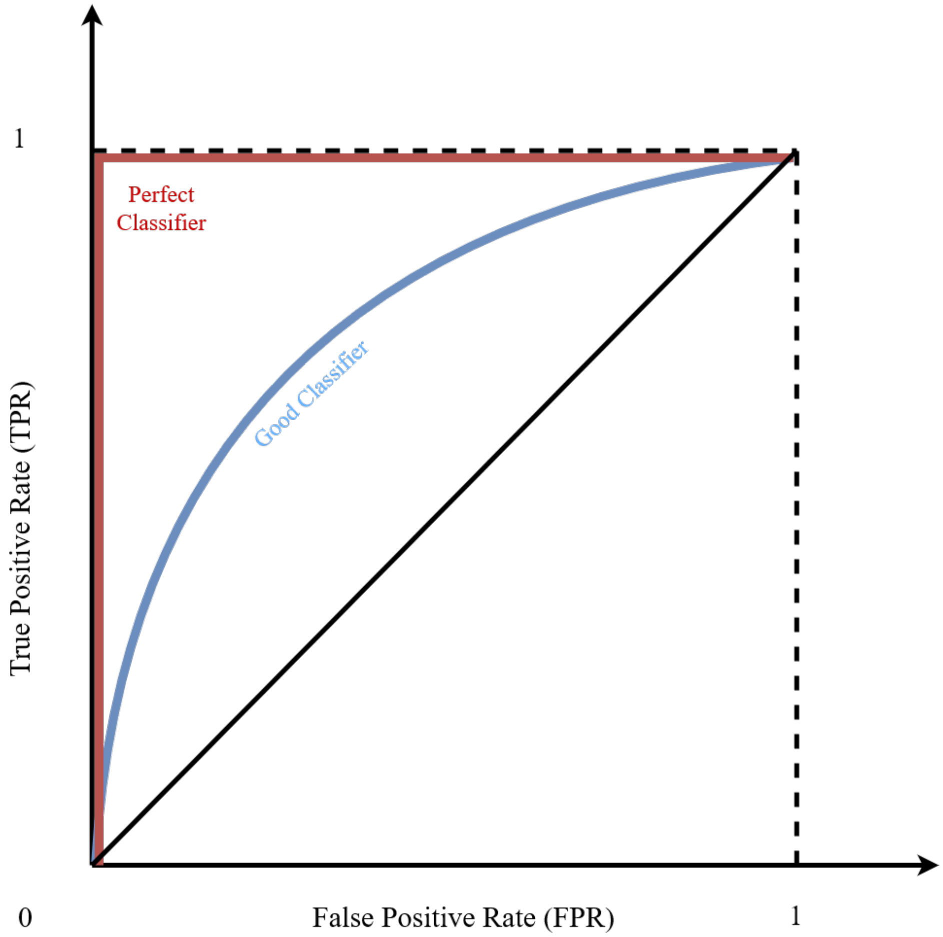


Fig. 1.7.1: A sample AUC graph with two different classifiers. A perfect classifier is close to the top left corner of the AUC graph.

1.8 Motivation

The motivation behind this research stems from the ever-growing importance of understanding cellular interactions within spatial transcriptomics data. While significant progress has been made in deciphering intracellular processes, we focus on unraveling the intricate web of cell-cell communication [6]. Recognizing that this facet plays a pivotal role in tissue function, development, and response to stimuli, there is a critical need for advanced computational models to navigate spatial data's complexities. Existing studies have laid a foundation, but challenges persist, such as dealing with spatial transcriptomics data's high dimensionality and sparsity [15]. The limited scope of previous models in capturing the nuanced relationships between diverse cell types and their interactions underscores the need for a more sophisticated approach.

Moreover, the shift in research emphasis from intracellular to intercellular interactions necessitates tailored methodologies. Our research seeks to address this gap by leveraging the power of machine learning, particularly Graph Neural Networks (GNNs) and Long Short-Term Memory (LSTM) networks [16]. This combination allows for a more accurate representation of spatial relationships and the sequential nature of cellular interactions. The spatial dimension adds an extra layer of complexity, and the need for effective feature selection methods is apparent. The integration of Backward Search provides a solution, enhancing the model's interpretability and performance.

Our motivation lies in advancing the understanding of cell-cell interactions within spatial transcriptomics data. By tackling the challenges unique to this domain, we aim to contribute valuable insights, laying the groundwork for more accurate predictions and opening avenues for broader applications in spatial transcriptomics and beyond.

1.9 Problem Statement

The research addresses the challenge of predicting cell-cell interactions within intricate spatial transcriptomics data, characterized by high dimensionality and complexity.

The spatial relationships inherent in the data are effectively encapsulated in a graph, $G = (V, E, X)$, where V represents individual cells, E signifies cell-cell interactions, and X incorporates attribute vectors for each node. The goal is to accurately predict the cell-cell communication. For this, we first employ ligand-receptor pairs extracted from spatial data to construct the graph G, where nodes denote cells and edges signify their interactions. This molecularly enriched graph serves as the basis for predicting cell-cell interactions within the spatial context.

To achieve this prediction, we combined a Graph Neural Network (GNN) with Long Short-Term Memor(LSTM), leveraging their respective proficiencies in handling graph-structured data and sequential information. The ligand-receptor pairs, integral to cellular interactions, are seamlessly integrated into the GNN framework. The resulting graph is represented by an adjacency matrix A, where a_{ij} is 1 if an edge exists between nodes v_i and v_j, and 0 otherwise. This research addresses the complexities of spatial transcriptomics by mathematically modeling cell-cell interactions, providing a robust framework for predictive analysis within high-dimensional spatial data.

1.10 Proposed Method

The objective is to predict cell-cell interactions within complex and high-dimensional spatial data. We have devised a pipeline tailored for this purpose, beginning with the conversion of spatial data into a graph format. This graph, denoted as $G = (V, E, X)$, encapsulates the spatial relationships inherent in the data. Here, V is a finite set of nodes representing individual cells, E is a set of edges signifying cell-cell interactions, and X encompasses attribute vectors associated with each node.

Crucially, the ligand-receptor pairs extracted from the spatial data contribute to the construction of this graph [13]. The molecular interactions between cells are captured within G, with nodes embodying individual cells and edges encapsulating the interactions between these cells. The attribute vector X linked to each node encapsulates pertinent information about the corresponding cell, allowing us to contextualize the spatial relationships within the data.

To carry out the prediction of cell-cell interactions, we employ a Graph Neural Network (GNN). This computational approach is particularly apt for handling graph-structured data. Furthermore, the ligand-receptor pairs, integral to cellular interactions, are effectively considered within this framework. The resulting graph is further represented by an adjacency matrix A, a square matrix of size $N \times N$ where N is the number of nodes. Elements a_{ij} in the adjacency matrix are defined as 1 if there exists an edge between nodes v_i and v_j, and 0 otherwise.

1.10.1 Contributions

Cell-cell communication has seen significant advancements in recent years, leading to a deeper understanding of the field. However, several challenges still need to be solved in dealing with the nature of spatial data, including sparsity and high dimensionality. This work addresses these challenges by introducing a new pipeline that combines graph neural networks with long short-term memory. This research builds upon previous work of GCNG[15] and SEAL [18], and some of the highlights of the research are listed below:

- We introduced an innovative pipeline for predicting cell-cell interactions. Our methodology is designed to handle various spatial transcriptomics datasets effectively, which can be challenging due to their high dimensionality and sparsity.

- We strategically combined Long Short-Term Memory (LSTM) with a graph neural network; our method goes beyond conventional approaches, offering a nuanced understanding of cellular relationships. It improved upon the existing methodology based on graph neural networks to produce the best results.

- We applied the proposed method to different spatial transcriptomics datasets. Additionally, we improved the existing methodologies (like GCNG), extending them by using other feature selection techniques to select the best features (genes). We provided a detailed outlook of our methodology by comparing it against the existing methods to highlight the improvement it provides.

- Leveraging the power of NetworkX and UMAP, we visually mapped cellular connectivity within spatial transcriptomics data. These visual insights contribute to a deeper understanding of the cellular landscape by highlighting different cell clusters with strong cell-cell interactions in the cell network.

References

[1] Bruce Alberts. *Molecular biology of the cell.* Garland science, 2017.

[2] Erick Armingol et al. “Deciphering cell–cell interactions and communication from gene expression”. In: *Nature Reviews Genetics* 22.2 (2021), pp. 71–88.

[3] Mary Ann Clark, Jung Choi, and Matthew Douglas. “Propagation of the Signal”. In: *Biology 2e* (2018).

[4] Jiawen Cui et al. “Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome”. In: *Plant Methods* 16 (2020), pp. 1–13.

[5] Ruben Dries et al. “Advances in spatial transcriptomic data analysis”. In: *Genome research* 31.10 (2021), pp. 1706–1718.

[6] Rui Hou et al. “Predicting cell-to-cell communication networks using NATMI”. In: *Nature communications* 11.1 (2020), p. 5011.

[7] Suoqin Jin et al. “Inference and analysis of cell-cell communication using CellChat”. In: *Nature communications* 12.1 (2021), p. 1088.

[8] Bogumił Kamiński et al. “A multi-purposed unsupervised framework for comparing embeddings of undirected and directed graphs”. In: *Network Science* 10.4 (2022), pp. 323–346.

[9] Zhaoyang Liu, Dongqing Sun, and Chenfei Wang. “Evaluation of cell-cell interaction methods by integrating single-cell RNA sequencing data with spatial information”. In: *Genome Biology* 23.1 (2022), pp. 1–38.

[10] Dimitrios Mathios et al. “Detection and characterization of lung cancer using cell-free DNA fragmentomes”. In: *Nature communications* 12.1 (2021), p. 5060.

[11] Sushmaa Chandralekha Selvakumar et al. “CRISPR/Cas9 and next generation sequencing in the personalized treatment of Cancer”. In: *Molecular Cancer* 21.1 (2022), p. 83.

[12] Kamilya Smagulova and Alex Pappachen James. “A survey on LSTM memristive neural network architectures and applications”. In: *The European Physical Journal Special Topics* 228.10 (2019), pp. 2313–2324.

[13] Aaron J Wilk et al. “Comparative analysis of cell–cell communication at single-cell resolution”. In: *Nature Biotechnology* (2023), pp. 1–14.

[14] Feng Xia et al. “Graph learning: A survey”. In: *IEEE Transactions on Artificial Intelligence* 2.2 (2021), pp. 109–127.

[15] Ye Yuan and Ziv Bar-Joseph. “GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data”. In: *Genome biology* 21.1 (2020), pp. 1–16.

[16] Yuansong Zeng et al. “Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks”. In: *Briefings in Bioinformatics* 23.5 (2022), bbac297.

[17] Li Zhang et al. “Identification of cell-to-cell interactions by ligand-receptor pairs in human fetal heart”. In: *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1866.12 (2020), p. 165917.

[18] Muhan Zhang and Yixin Chen. “Link prediction based on graph neural networks”. In: *Advances in neural information processing systems* 31 (2018).

CHAPTER 2

Prediction of Cell-cell Communication from Spatial Transcriptomics Data Using a Long Short-term Memory Graph Neural Network

2.1 Introduction

Cell-cell interactions form a complex network of signaling pathways where molecular signals intricately regulate cellular activities . This intricate interplay is essential for orchestrating diverse physiological functions, from development and tissue homeostasis to immune responses [20]. Cells communicate through a complex network of signalling pathways, where molecular signals are exchanged to regulate cellular activities [23]. Understanding cell-cell interactions is crucial for deciphering the mechanisms underlying health and disease. Recent advancements in technologies like single-cell RNA sequencing and spatial transcriptomics have provided unprecedented insights into the molecular dialogues between cells, enabling the exploration of intra- and intercellular interactions. In contemporary research, there has been a notable transition in focus from intracellular interactions, which pertain to molecular processes occurring within individual cells, to intercellular interactions [24], which involve communication and signaling between neighboring cells. This shift reflects a growing recognition of the significance of cell-cell communication in the intricate orchestration of biological processes, shedding light on the dynamic interplay between cells within complex biological systems. [2]. This is where spatial data comes into play.

In the context of biological research, spatial data refers to information about the spatial arrangement and distribution of biological entities, such as cells within tissues and organs [5]. Unlike traditional gene expression data, spatial data capture the specific location of individual cells, providing a spatial context to molecular information [4]. This type of data is crucial because it offers a three-dimensional perspective on the cellular landscape, allowing researchers to explore how cells are organized in tissues. The importance of spatial data lies in its ability to unravel the spatial relationships between cells, shedding light on the intricate architecture of tissues. By understanding the spatial context of gene expression, researchers can decipher the spatial organization of cell types, identify signalling gradients, and investigate how cells interact within their microenvironment [17]. This spatial awareness is crucial for unraveling the complexities of cellular behavior.

2.2 Related Work

In line with this, advanced computational tools, such as Single-cell graph neural network (scGNN) [19], have emerged as invaluable instruments for analyzing single-cell RNA sequencing (scRNA-seq) data. Key features encompass modeling cell-type-specific regulatory signals and formulating cell relationships from a pruned Graph Neural Network (GNN) cell graph. The GNN cell graph is a structural foundation, representing cellular interactions and regulatory influences. The graph is refined through pruning to emphasize essential connections, facilitating a more focused analysis of the intricate web of cellular relationships and regulatory signals specific to individual cell types. scGNN demonstrated superior performance over existing tools. Integrating gene regulatory signals and cell network representations contributes to its success, with promising results validated across various datasets, showcasing its potential in unraveling intricate cell relationships and contributing to disease studies. Addressing the challenges of high dimensionality and sparsity in RNA-sequencing data, the paper offers insights into how the proposed framework navigates these issues, enhancing its applicability to real-world datasets and bolstering its robustness in capturing meaningful biological information within vast and sparse genomic landscapes [19].

DeepSignalingLinkNet [11] is a deep-learning model that predicts signaling cascades critical in cancer biology by integrating transcriptomics and copy number data with protein-protein interactions. Addressing the limitations of existing models that rely on gene co-expression networks or shortest-path-based protein-protein interaction analyses, deepSignalingLinkNet is designed to predict direct and sparse signaling (Predicting specific signaling cascades with a limited molecular set.) cascades precisely. It ensuring a more focused and targeted approach rather than encompassing the entire network. This precision enhances the utility of deepSignalingLinkNet in uncovering key regulatory pathways without unnecessary complexity, contributing to a more nuanced understanding of cellular signaling events. Trained on curated KEGG signaling pathways, the model identifies informative omics and protein-protein interaction features in a data-driven manner. The broader context underscores the significance of understanding signaling pathways in cancer molecular biology and the challenges in uncovering comprehensive signaling networks that regulate tumor development and drug response [11].

Prior research has mainly focused on deciphering intracellular interactions, revealing critical challenges within the broader context of cell-cell interactions. These investigations, while valuable, underscore the need for a more comprehensive exploration of intercellular dynamics to capture the broader landscape of cellular communication within the spatial context. The transition from intracellular to intercellular studies represents a critical shift in focus, recognizing the intricate interplay between cells and emphasizing the importance of studying cell interactions within their spatial microenvironment.

The graph convolutional neural networks for inferring gene interaction (GCNG)

[25] is a novel method to infer gene interactions involved in cell-cell communication from spatial single-cell expression data. It leverages graph convolutional neural networks (GCNs) to encode the spatial relationships between cells and the expression of gene pairs within each cell. Unlike traditional correlation-based methods, GCNG captures intra- and inter-cellular interactions using the graph structure and gene expression data. The method is particularly suitable for spatial transcriptomics, where high-throughput information about intra- and inter-cellular interactions is available. GCNG outperforms unsupervised methods such as Giotto and Spatial PC, providing accurate predictions of extracellular gene interactions involved in cell-cell communication. Its effectiveness is demonstrated on SeqFISH+ and MERFISH datasets, showcasing its ability to predict meaningful interactions and overcome the limitations of unsupervised methods [25].

Hist2ST [27] represents a groundbreaking deep-learning paradigm tailored to predict RNA-seq gene expression within the spatial transcriptomics framework, leveraging histology images. This innovative model adopts a multifaceted approach involving extracting 2D visual features, capturing details such as shapes and spatial relationships through convolutional modules, incorporating transformer and graph neural network modules to capture spatial relations, and, notably, integrating a Long Short-Term Memory (LSTM) mechanism. LSTM, as a recurrent neural network, introduces a temporal learning dimension to the model, enabling it to discern and utilize temporal dependencies within the data effectively. Through extensive testing on diverse cancer and standard datasets, Hist2ST emerges as a frontrunner, demonstrating superior gene expression prediction and spatial region identification capabilities compared to existing methodologies. Furthermore, pathway analyses highlight Hist2ST's proficiency in preserving critical biological information. It solidifies its standing as a robust tool for extracting spatial transcriptomics insights from histology images and unraveling intricate tissue molecular signatures [27].

Both of these papers highlight the importance of spatial data, but they still need to employ appropriate feature selection methodologies. The significance of feature selection lies in its ability to enhance model performance, reduce overfitting, and enhance interpretability. In many datasets, redundant or irrelevant features often introduce noise and decrease the efficiency of a model. Additionally, in high-dimensional data, selecting a subset of relevant genes not only improves model efficiency but also helps reduce computational costs [26].

Feature selection techniques play a crucial role in enhancing the predictive power of models, especially in complex tasks like predicting cell-cell interactions from spatial transcriptomics data. In this research, three distinctive feature selection methods, backward search, information gain, and chi-squared, were strategically chosen for their unique strengths and relevance to the study's objectives. While various methods exist, our selection was motivated by their proven effectiveness in similar studies and their suitability for capturing spatial transcriptomics patterns [9].

Backward search stands out as an effective technique for feature selection due to its systematic approach of iteratively removing features to improve model performance. This method optimizes the model's capacity to capture relevant information while mitigating the risk of overfitting, contributing to the robustness of the predictions [7].

Information gain is a widely used method in machine learning and provides a principled way to assess the relevance of features based on their ability to reduce uncertainty or entropy in the dataset. Its capability to measure the impact of features on the overall dataset's entropy aligns well with the study's objective of uncovering significant contributors to cell-cell interactions [2].

Chi-squared is a statistical test commonly used for categorical data which complements the feature selection process by evaluating genes' independence and association with cell-cell interactions. In spatial transcriptomics, where genes often exhibit complex relationships, chi-squared analysis provides a statistical lens to discern the significance of each gene in contributing to the observed interactions. By considering the distribution of genes and their associations with cell-cell interactions, chi-squared augments the feature selection process, offering insights into genes that might play pivotal roles in mediating spatial cellular communication [16].

Filter Methods: In exploring feature selection, metrics like information gain, chi-squared, and correlation coefficients are computationally efficient but may overlook feature dependencies. It motivates our strategic choice of methods to balance efficiency and nuanced relationship capture in spatial transcriptomics data.

Wrapper Methods: As an alternative, wrapper methods, like Recursive Feature Elimination (RFE) and forward/backward search, involve iterative model training to assess feature subsets' impact on performance. This nuanced approach aligns with our selection rationale, balancing efficacy and computational efficiency in the context of spatial transcriptomics data. [22].

Each technique has advantages and trade-offs, making the selection dependent on the dataset characteristics and modeling goals. For instance, filter methods are computationally efficient but may not capture feature dependencies, while wrapper methods can be more accurate but computationally intensive. The choice depends on the dataset size, computational resources, and the interpretability required for the specific task [22]. By enhancing predictive model performance and offering insights into the molecular nuances of spatial transcriptomics data, these techniques play a pivotal role in unraveling complex cellular interactions. Our research focuses on deciphering cell connections, where nodes symbolize individual cells and edges represent cell-cell interactions. To predict these interactions, we employed a methodology akin to the GCNG approach.

2.3 Materials and Methods

2.3.1 Datasets

Exploring the spatial nuances of gene expression, our study delves into publicly available datasets, namely seqFISH+ and MERFISH. These spatial transcriptomics datasets capture the intricate spatial organization of gene expression within tissues, offering a distinctive perspective. seqFISH+ captures gene expression patterns in intact tissues, enabling the exploration of intricate cellular relationships. On the other hand, MERFISH employs a combinatorial labelling strategy, facilitating the high-throughput mapping of RNA transcripts with spatial precision. They form the foundation for investigations into cellular heterogeneity, spatial organization, and gene expression dynamics within complex biological environments. In this study, both datasets are from the mouse cortex tissue; seqFISH+ comprises 3,000 gens for 10,000 cells, while MERFISHis consists of 10,050 cells with 1,368 genes.

Table 2.3.1 depicts the details of datasets including the number of cells, number of genes and tissue.

Table 2.3.1: Details of the datasets used in this work.

Dataset	Accession	Tissue	No. of Cells	No. of Genes
seqFISH+		Mouse Cortex	10,000	3,000
MERFISH	GSE202638	Mouse Cortex	10,050	1,368

2.4 Proposed Method

The objective is to predict cell-cell interactions within complex and high-dimensional spatial data. We have devised a pipeline tailored for this purpose, beginning with the conversion of spatial data into a graph format. This graph, denoted as $G = (V, E, X)$, encapsulates the spatial relationships inherent in the data. Here, V is a finite set of nodes representing individual cells, E is a set of edges signifying cell-cell interactions, and X encompasses attribute vectors associated with each node.

Crucially, the ligand-receptor pairs extracted from the spatial data contribute to the construction of this graph [21]. The molecular interactions between cells are captured within G, with nodes embodying individual cells and edges encapsulating the interactions between these cells. The attribute vector X linked to each node encapsulates pertinent information about the corresponding cell, allowing us to contextualize the spatial relationships within the data.

To carry out the prediction of cell-cell interactions, we employ a Graph Neural Network (GNN). This computational approach is particularly apt for handling graph-structured data. Furthermore, the ligand-receptor pairs, integral to cellular interactions, are effectively considered within this framework. The resulting graph is further represented by an adjacency matrix A, a square matrix of size $N \times N$ where N is the number of nodes. Elements a_{ij} in the adjacency matrix are defined as 1 if there exists an edge between nodes v_i and v_j, and 0 otherwise.

The proposed methodology based on a modified version of the SEAL (learning from Subgraphs, Embeddings and Attributes for Link prediction) method [28] consists of three steps, namely, 1) Preprocessing step, 2) Network Construction, and 3) Prediction. The preprocessed data is used to construct the graph network based on which the graph neural network makes the prediction. The fundamental steps comprise certain substeps, including feature selection, subgraph extraction, construction of node information matrix, and learning from the graph convolution neural network. These steps will be explained in the upcoming section of the paper and are shown in figure 2.4.1.

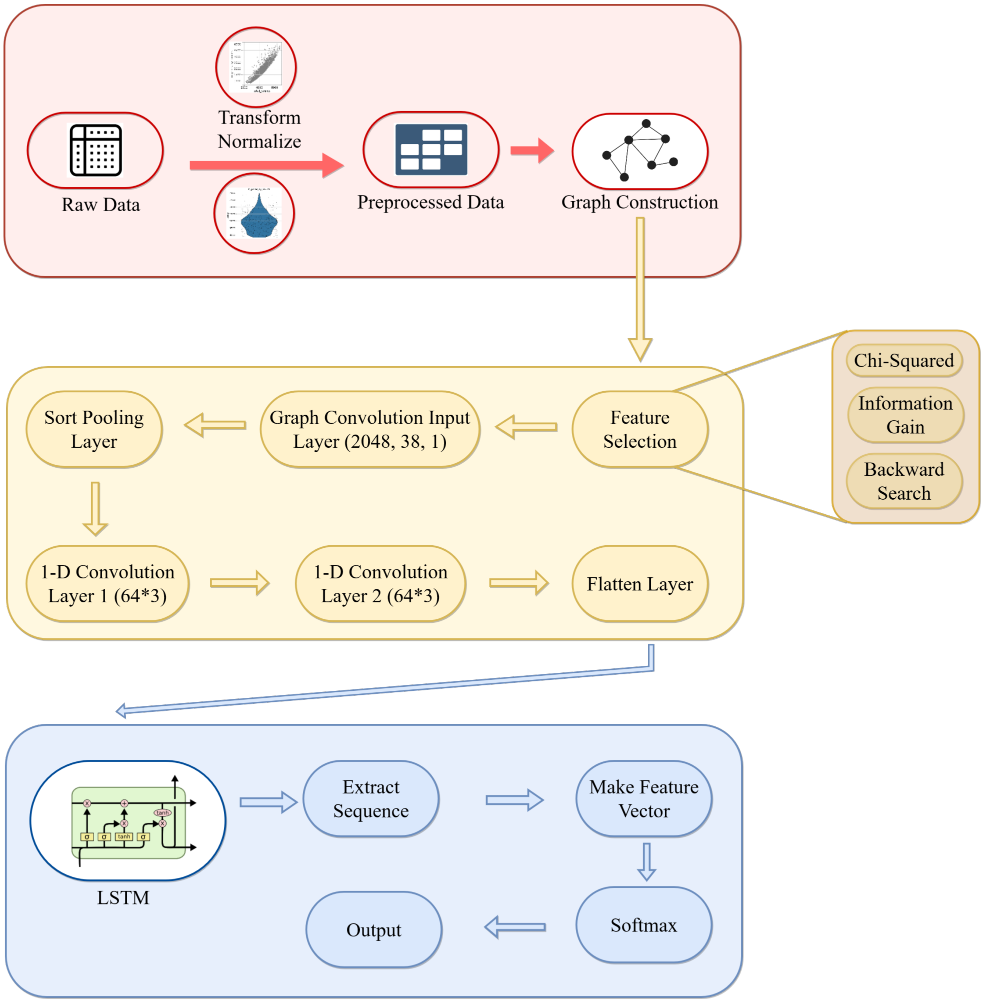


Fig. 2.4.1: The overall proposed methodology with various steps involved.

2.4.1 Step 1: Data Preprocessing

Data preprocessing is crucial to render the data suitable for analytical pursuits. The normalization process is integral, involving the scaling of individual samples to a consistent magnitude. This could entail adjusting gene expression values within cells in spatial data, ensuring a uniform scale across varied genes. For instance, normalizing expression levels within each cell to a standardized sum value facilitates comparability. Simultaneously, transformation techniques may address challenges like high dimensionality or skewed gene expression distributions [13]. Common transformations, such as logarithmic conversions, mitigate skewed distributions, enhancing the dataset's adaptability to specific analyses. Leveraging the power of Scanpy, a versatile Python library, we conducted robust data preprocessing tasks, including transformation and normalization. Scanpy's capabilities, inspired by Seurat, ensure the generation of high-quality, standardized datasets, providing a solid foundation for subsequent analyses [18].

Additionally, given the inherent sparsity in spatial data, employing sparse matrix representations or specialized algorithms tailored for sparse datasets becomes pertinent. Spatial coordinate transformations might also be employed, particularly when integrating data from disparate spatial sources or aligning diverse spatial datasets to a standard coordinate system. The amalgamation of these preprocessing strategies is instrumental in refining data quality, reducing noise, and establishing a conducive foundation for meaningful spatial analyses and modeling. [19]

2.4.2 Step 2: Network Construction

Determining the neighbors of each cell involves a spatial approach using Euclidean distance calculations in the image coordinates. The Euclidean distance is computed for all cell pairs, and a distance threshold is applied to identify neighboring cells. The selection of the threshold is a crucial step, and in this study, it was determined through 10-fold cross-validation. For 2D images, the threshold value was chosen to represent the number of neighbors in physical contact with a cell. Taking the seqFISH+ cortex data as an example, the set of neighbors was used to construct an adjacency matrix (A) with a size of 10,000 x 10,000, where 10000 is the number of cells in the seqFISH+ dataset. The binary adjacency matrix $A^{(b)}$ is symmetric, with $A^{(b)}_{ij} = A^{(b)}_{ji} = 1$ if cells i and j are neighbors and 0 otherwise. Similarly, a weighted adjacency matrix $A^{(w)}$, with $A^{(w)}_{ij} = A^{(w)}_{ji} =$ weight (euclidean distance) between cells i and j if cells i and j are neighbors, and 0 otherwise. The two matrices are combined to obtain a whole adjacency matrix, which is then normalized. This matrix serves as a representation of the spatial relationships between cells in the dataset. The seqFISH+ dataset is divided into seven "fields of view". Each field of view represents a distinct portion of the sample that is imaged or analyzed separately. In the SeqFish+ dataset, the profiling includes information from seven fields of view. These fields of view have been selected strategically to capture a comprehensive understanding of the biological sample, and analyzing multiple fields of view provides a more nuanced view of the spatial gene expression patterns within the tissue or cells under investigation [25].

2.4.3 Step 3: Prediction

After constructing a graph network, the next step is link prediction. It starts with feature selection. Later, the prediction phase involved constructing a comprehensive node information matrix comprising node embeddings, labels, and attributes. This matrix was then fed into a Graph Neural Network, complemented by LSTM, to predict intricate cell-cell interactions in spatial transcriptomics datasets.

2.4.3.1 Feature/Gene Selection

Feature selection is a critical component in the preprocessing pipeline, focused on identifying and retaining the most informative features while discarding irrelevant or redundant ones [1]. Information gain, chi-squared, and backward search are methods that we choose to achieve this objective. These three feature selection techniques are widely utilized. A comprehensive comparison is provided, emphasizing the advantages of using filter and wrapper methods. Unlike some modern feature selection techniques, these techniques use less computation power [1]. These techniques enhance the efficiency and effectiveness of subsequent analyses, ensuring that the model is built on a refined set of features that genuinely capture the underlying patterns and relationships in the spatial data.

Chi-squared The chi-squared test is a statistical method used to determine if there is a significant association between two categorical variables. It compares the expected and observed frequencies in a contingency table. The resulting chi-squared statistic helps determine the significance of the association.

The chi-Squared (χ^2) test is a statistical method to assess the independence or dependence between categorical variables within a dataset. Originating from inferential statistics, the chi-Squared test operates on data that can be categorized into distinct groups. It is particularly useful when dealing with nominal data, where observations fall into discrete categories but lack a natural order.

The fundamental premise of the chi-Squared test is to compare the observed frequencies of occurrences in a contingency table with the frequencies that would be expected if the variables under investigation were independent. The test calculates a statistic (χ^2) based on the squared differences between observed and expected frequencies, normalized by the expected frequencies. The larger the resulting χ^2 value, the greater the discrepancy between observed and expected values. Subsequently, the calculated χ^2 value is compared to a critical value from the chi-Squared distribution to determine whether the observed and expected frequencies significantly deviate from independence. If the calculated χ^2 value exceeds the critical value, it indicates a significant association between the variables. Conversely, a lower χ^2 value suggests independence.

The chi-squared statistic is calculated as follows:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- χ^2 is the chi-squared statistic.
- O_i is the observed frequency in each category.
- E_i is the expected frequency in each category [22].

Information Gain Information gain is a metric used in feature selection to assess the significance of a feature in terms of its ability to contribute relevant information to a predictive model. Particularly employed in decision trees and machine learning, information gain quantifies the reduction in uncertainty about the target variable achieved by considering a specific feature.

The calculation of information gain involves evaluating the entropy or impurity of a dataset before and after considering a feature. Entropy represents the measure of disorder or unpredictability in a set of data. A high information gain implies that a feature provides substantial information about the target variable, making it crucial for effective prediction. In essence, information gain aids in selecting features that enhance the model's predictive performance by reducing uncertainty, thereby contributing to more accurate and efficient machine learning models.

It quantifies the reduction in entropy (uncertainty) achieved by dividing the dataset based on the values of a particular variable. Information gain ranks features based on how well they contribute to the classification task [22].

$$\text{IG} = \text{Entropy}_{\text{original}} - \sum \left(\frac{|\text{Subset}|}{|\text{Total}|} \times \text{Entropy}_{\text{subset}} \right)$$

Backward Search Backward search, a feature selection method, operates by iteratively removing features from a model to enhance its performance. This method begins by including all available features and systematically eliminates those deemed less relevant or redundant. The process is driven by the model's performance, which is assessed at each step of feature removal. In machine learning, especially when dealing with high-dimensional datasets, backward search is valuable in enhancing model efficiency, reducing overfitting, and improving interpretability. The method aims to find the optimal subset of features that maximizes predictive accuracy while minimizing computational complexity. The backward search algorithm often employs a criterion such as accuracy, AUC (Area Under the Curve), or other relevant metrics to evaluate the model's performance after removing each feature. Features that contribute less to the model's predictive power are successively pruned, leading to a refined set that retains only the most informative attributes. It also has a memory mechanism to store subsets of features that have been evaluated to avoid redundant evaluations.

While backward search is effective, its computational cost may be relatively high, especially when dealing with a large number of features. Therefore, it is essential to strike a balance between feature reduction and computational efficiency, considering the specific characteristics of the dataset and the goals of the machine learning task at hand. Overall, backward Search is a valuable tool in feature selection methods, offering a systematic and data-driven approach to optimizing model performance [8].

The steps carried out in backward search are defined in algorithm 2.4.1.

Algorithm 2.4.1 Backward Search with Memory

```
Initialize: Start with the entire set of features: CurrentFeatures = AllFeatures.
Iteration:
while stopping criterion is not met do
    for all features in CurrentFeatures do
        Remove the feature and form a new subset: NewSubset = CurrentFeatures - {FeatureToRemove}.
        if NewSubset is not in the Memory then
            Evaluate the performance criterion on the model with features in NewSubset.
            Update the Memory with NewSubset.
            if the performance improves then
                Update CurrentFeatures to be NewSubset.
                Update the selected features if needed.
Output: Return the final selected features.
```

2.4.3.2 Subgraph Extraction

Based on the SEAL method, we extracted the enclosing subgraph. Enclosed subgraph extraction is a process that involves capturing the local neighborhood structures around specified target nodes, denoted as u and v. In this context, "enclosing subgraph" refers to the subgraph induced by considering the union of neighbors of both u and v up to a certain number of hops (k-hops) [10] in the network. This subgraph is tailored for a node pair (u, v). The process is crucial for link prediction tasks, where the objective is to predict the existence or absence of links between nodes. The training data contains both positive (existing) and negative (non-existent) links, sampled based on h-hop neighbors for the target nodes u and v. It involves examining the local structures up to a specified distance around the nodes of interest. Utilizing subgraphs enables the calculation of various first-order heuristics, like familiar neighbors, which are valuable features for predicting links in the network. [28].

2.4.3.3 Node Information Matrix Construction

This step's central part is constructing Node Information Matrix, X, which represents the features or attributes associated with each node (cell) in the spatial graph. It has three subcomponents including node labelling, node embedding and node attribute we explain them in the following paragraphs.

Node Labelling: The Double-Radius Node Labeling (DRNL) algorithm is devised for assigning labels to nodes in enclosing subgraphs precisely to distinguish target nodes x and y in the context of link prediction. The labels are assigned based on the double-radius of nodes concerning the target nodes, represented as $(d(i,\ x),\ d(i,\ y))$. The algorithm starts by assigning label 1 to the target nodes x and y. Then, labels are assigned to other nodes based on their double-radius values. DRNL ensures that nodes with similar relative positions receive similar labels, and it achieves benefits where the magnitude of labels reflects the distance to the center. The algorithm's distinctive features include satisfying specific criteria for nodes i and j based on their distances to target nodes, ensuring the preservation of magnitude information in labels. The DRNL algorithm is handy for training or ranking nodes when node labels are used. The node labeling includes a perfect hashing property, allowing for closed-form computation, and a lookup table for DRNL is presented for practical implementation. Notably, the calculation of distances involves temporarily removing the influence of the other target node to capture the distance between nodes. The algorithm works to satisfy the following conditions:

1. if $d(i,x) + d(i,y) \neq d(j,x) + d(j,y)$, then $d(i,x) + d(i,y) < d(j,x) + d(j,y) \Leftrightarrow f_l(i) < f_l(j)$;

2. if $d(i,x) + d(i,y) = d(j,x) + d(j,y)$, then $d(i,x)d(i,y) < d(j,x)d(j,y) \Leftrightarrow f_l(i) < f_l(j)$.

where $f_l(i)$ is the label assigned to node i and $(d(i,x), d(i,y))$ is the double radius [28].

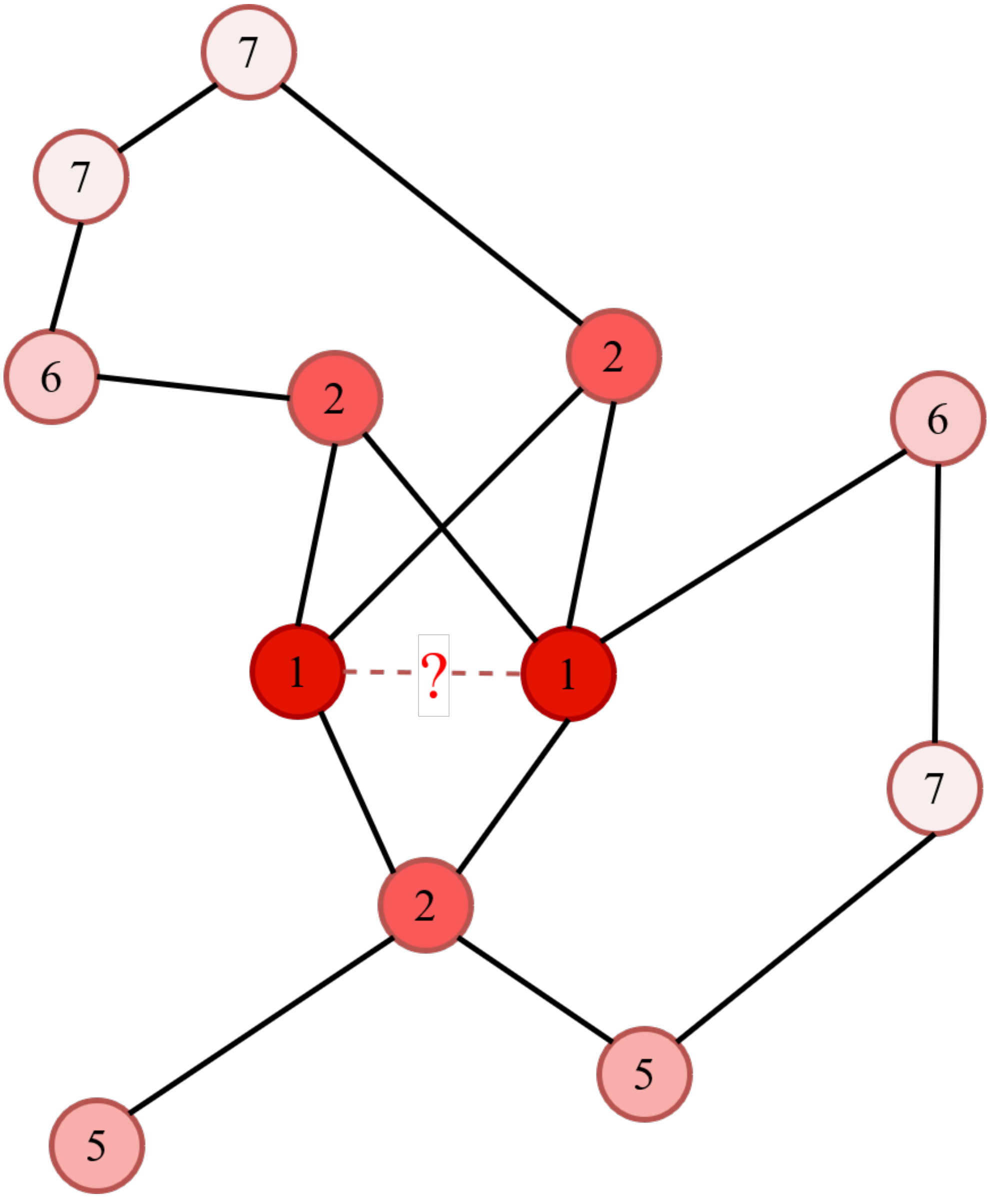


Fig. 2.4.2: The labelling of the nodes based on their distance from the target node.

As illustrated in figure 2.4.2, Double-radius node labeling calculates two types of distances for each node: the shortest path distance to other nodes and the distance to nodes within a specified radius. The first radius encompasses nodes reachable within a single step, while the second radius extends the reach to nodes reachable within two steps. Combining these two distance metrics results in a label for each node, reflecting its local and slightly more distant neighborhood relationships. This dual-radius labeling approach is particularly useful in graph-based applications where capturing local and semi-local connectivity patterns is crucial for graph classification, clustering, or link prediction tasks. The resulting node labels serve as feature representations that encode information about the node's immediate surroundings and broader network context.

Node Attributes: For the prediction of a link, three primary feature types are utilized. Inherent in observed node and edge structures, graph structure features encompass link prediction heuristics, centrality scores, and structural patterns. Latent features, obtained through matrix factorization, provide low-dimensional node representations, emphasizing global properties but facing challenges in capturing structural similarities. Explicit features, given as attribute vectors, encompass any non-structural side information. Examples include word distributions in citation networks and user profiles in social networks. While graph structure features are handcrafted and limited, latent features may struggle with interpretability, and explicit features offer additional context beyond network structure [28].

Node Embeddings: The challenge arises when directly generating embeddings on the observed network G=(V,E) using positive and negative training links. If embeddings are generated directly on G, the model may focus too much on fitting the link existence information of the training links, leading to poor generalization performance. A technique known as *"negative injection"* is employed to address the issue at hand. This involves temporarily incorporating the sampled negative training links (E_n) into the set of edges (E), thereby creating a new graph $G' = (V, E \cup E_n)$. By doing so, both positive and negative training links share the same link existence information in the embeddings. This strategic move prevents the model from solely optimizing for the link existence information, leading to improved generalization performance. The negative injection trick aims to enhance SEAL's learning process by mitigating the bias introduced by focusing solely on link existence during training [28].

2.4.3.4 Graph Neural Network

Following this, the graph convolutions are employed to extract hidden feature information from nodes in a graph. The network comprises four graph convolutional layers, each followed by a hyperbolic tangent (tanh) non-activation function. Tanh is an activation function that squashes the output to the range [-1, 1]. The architecture involves stacking these convolutional layers to extract multi-hop node features. This stacking enables the model to capture information from multiple graph levels, considering different numbers of hops from each node. Additionally, the node states from each layer are concatenated to form the final node states. This mechanism enhances the model's ability to capture and integrate information across various local substructures, contributing to the overall predictive performance of the network [28].

The convolutional and dense layers play a crucial role in making predictions based on the sorted graph representations generated by the SortPooling layer. The network takes as input the adjacency matrix A of a graph G with n nodes, where each node has a c-dimensional feature vector represented by the node information matrix X of an enclosing subgraph.

Underline SEAL's DGCNN architecture comprises of adjacency matrix $A \in \{1, 0\}^{nXn}$ of graph G with n number of nodes and each node containing the c dimensional feature vector as well as the node information matrix $X \in R^{nXc}$ of an enclosing subgraph with each row representing the node, DGCNN employs the following convolution layer:

$$Z = f(D\tilde{-}1\tilde{A}XW)\,, \tag{1}$$

where $\tilde{A} = A + I$, I is the identity matrix, $\tilde{D}$ is the diagonal degree matrix with $\tilde{D_{i,i}} = \sum j\tilde{A_{i,j}}$, W is a trainable graph convolutional parameters, f is a non-linear activation function, and $Z \in R^{nXc'}$ is the output activation matrix [28].

This convolutional operation allows the model to learn and capture complex patterns and features within the graph structure, contributing to the overall predictive capabilities of the network [28].

2.4.3.5 Long Short-Term Memory Layer

The addition of an LSTM layer (Long Short-Term Memory) to the SEAL pipeline brings about many benefits, including noise reduction by filtering out irrelevant information and focusing on relevant sequences, improved generalization by considering the sequential nature of the data by allowing the model to discern patterns and context across sequential information, LSTM's can improve the model's ability to generalize to unseen data points while reducing noise.[14]. These improvements are shown in the later section of the paper when our model is compared against the SEAL. These results are visible in table 2.6.1 and table 2.6.2.

2.5 Performance Evaluation

Our model's effectiveness was assessed using the Area Under Curve (AUC) and Receiver Operating Characteristic (ROC) curve as metrics. The choice of AUC and ROC is pivotal in binary classification scenarios, particularly in predicting interactions. The ROC curve, fundamental to AUC computation, charts the trade-off between true positive and false positive rates across diverse threshold settings. Opting for AUC is strategic as it comprehensively evaluates the model's ability to discriminate between positive and negative instances, transcending the influence of a specific threshold setting. This choice is rooted in the robustness and versatility of AUC and ROC metrics, making them well-suited for the nuanced evaluation required in predicting cell-cell interactions. [28].

This involved using both training and testing datasets, encompassing positive (existent) and negative (non-existent) links. The negative set was created by randomly selecting an equal number of unconnected pairs of nodes from the network where no edge connection exists. Notably, these negative pairs were ligand-receptor pairs chosen randomly from non-interacting gene data. On the other hand, the positive set comprised known interactions between the samples. These are pairs of genes known to interact in the given sample. To create the negative set, we randomly selected ligand-receptor pairs that are not known to interact among the sample. For the training and testing split, 10% of the links were arbitrarily removed and set aside as testing data, while the remaining 90% were utilized as training data. This division was crucial for assessing the model's generalization and predictive capabilities on unseen data [25].

SEAL's γ decaying theory highlighted that a small number of hops is enough to extract high-order heuristics that help focus on local subgraph structures to capture complex patterns and relationships. SEAL suggests that such patterns can be effectively captured within a small neighborhood around the target nodes. This concept is essential for computation efficiency and also aligns with the observation that local structures often provide valuable information for predicting links in a network [28].

We used three popular feature selection techniques : chi-squared, information gain, and backward search [22].

The number of features in a machine learning model is a critical decision involving a trade-off between model complexity and generalization performance. We chose 70 features for all the three feature selection techniques. Increasing the number of features improves the performance to a certain extent. However, it also gives rise to other challenges like the risk of overfitting, where the model performs well on the training data but fails to generalize to new, unseen data, and increased Complexity, where more features can introduce noise and Complexity into the model [6]. This will give us a healthy balance between the model complexity and performance.

Based on these performance metrics and feature selection techniques, we compared our model against two other approaches to get a more comprehensive outlook.

One of these approaches is GCNG; the Graph Convolutional Neural Networks for Genes (GCNG) is a novel approach designed to analyze spatial transcriptomics data. In spatial transcriptomics, the goal is to understand the gene expression patterns of the spatial organization of cells. GCNG leverages the power of graph convolutional neural networks (GCNs) to integrate spatial information and gene expression data for a more comprehensive analysis [25].

The other one is SEAL. SEAL is a methodology developed for the task of link prediction in networks. Link prediction involves anticipating or predicting connections between nodes in a network, and it has applications in various domains, such as social networks, biological networks, and recommendation systems. SEAL introduces several innovative concepts to enhance the accuracy of link prediction [28].

This also gives us a chance to compare the result of SEAL and our modified version of SEAL, making the results apparent and highlighting the advantage of using the LSTM layer.

2.6 Results and Discussion

Our results presents a comprehensive evaluation of three distinct feature selection methods, namely chi-Squared, information gain, and backward search, employing three cutting-edge approaches—LSTM-SEAL, SEAL, and GCNG—on two spatial transcriptomics datasets, SeqFISH+ and MERFISH. The results are systematically summarized in two tables, encapsulating the Area Under the Curve (AUC) values for various feature selection methods and approach combinations. Table 2.6.1 delineates the performance metrics for the SeqFISH+ dataset, while Table 2.6.2 focuses on the MERFISH dataset.

Moreover, two AUROC graphs visually represent the true positive rate against the false positive rate.

The construction of the AUC graph involves plotting the ROC (Receiver Operating Characteristic) curve, where the true positive rate is plotted against the false positive rate across different threshold values. This curve visually encapsulates the model's discrimination capabilities across threshold settings. The steeper the ROC curve, the more adept the model is at distinguishing between positive and negative classes. The AUC itself is calculated as the area under this curve, with a perfect model achieving an AUC of 1 and a model with no discriminatory power yielding an AUC of 0.5 [28]. Interpretation of the AUC graph involves an analysis of its shape and steepness, providing insights into the model's overall performance. A model that consistently achieves higher true positive rates across a range of false positive rates is deemed superior. The point (0,1) on the graph signifies perfect sensitivity and specificity. Ultimately, the AUC graph is a comprehensive visualization aiding in comparing and

Table 2.6.1: Comparison of our method with other approaches for seqFISH+ dataset based on AUC values.

Selection Method	Approach	10 Features	20 Features	30 Features	40 Features	50 Features	60 Features	70 Features
Chi-Squared	LSTM-SEAL	0.4621	0.5239	0.5569	0.5998	0.6981	0.7189	0.7334
	SEAL	0.4019	0.5412	0.5562	0.5834	0.6655	0.6997	0.7119
	GCNG	0.4129	0.5389	0.5441	0.6089	0.6794	0.7245	0.7387
Information Gain	LSTM-SEAL	0.4556	0.6013	0.6318	0.6589	0.7048	0.7234	0.7389
	SEAL	0.4489	0.5689	0.6019	0.6455	0.6858	0.6987	0.7067
	GCNG	0.4626	0.5833	0.6159	0.6419	0.6446	0.6768	0.6910
Backward Search	LSTM-SEAL	**0.5434**	**0.6429**	**0.7259**	**0.7566**	**0.8366**	**0.9119**	**0.9281**
	SEAL	0.4811	0.6334	0.6778	0.7081	0.7969	0.8876	0.8903
	GCNG	0.4511	0.6023	0.6589	0.6832	0.7922	0.8345	0.8415

evaluating different models, enabling practitioners to make informed decisions about model selection based on their discriminatory power [3].

Table 2.6.2: Comparison of our method with other approaches for MERFISH dataset based on AUC values.

Selection Method	Approach	10 Features	20 Features	30 Features	40 Features	50 Features	60 Features	70 Features
Chi-Squared	LSTM-SEAL	0.4716	0.6001	0.6259	0.6311	0.6778	0.7081	0.7203
	SEAL	0.4589	0.5861	0.6008	0.6189	0.6482	0.6834	0.6890
	GCNG	0.3926	0.5693	0.5771	0.5912	0.6221	0.6511	0.6522
Information Gain	LSTM-SEAL	0.4212	0.5728	0.6063	0.6189	0.6784	0.7169	0.7089
	SEAL	0.4026	0.5589	0.5911	0.6260	0.6588	0.6987	0.7068
	GCNG	0.3605	0.5256	0.5598	0.5789	0.6239	0.6456	0.6672
Backward Search	LSTM-SEAL	**0.4833**	**0.6332**	**0.6988**	**0.7345**	**0.8256**	**0.8891**	**0.9129**
	SEAL	0.4758	0.6005	0.6711	0.6978	0.7986	0.8674	0.8798
	GCNG	0.4053	0.5436	0.6333	0.6544	0.7558	0.8098	0.8245

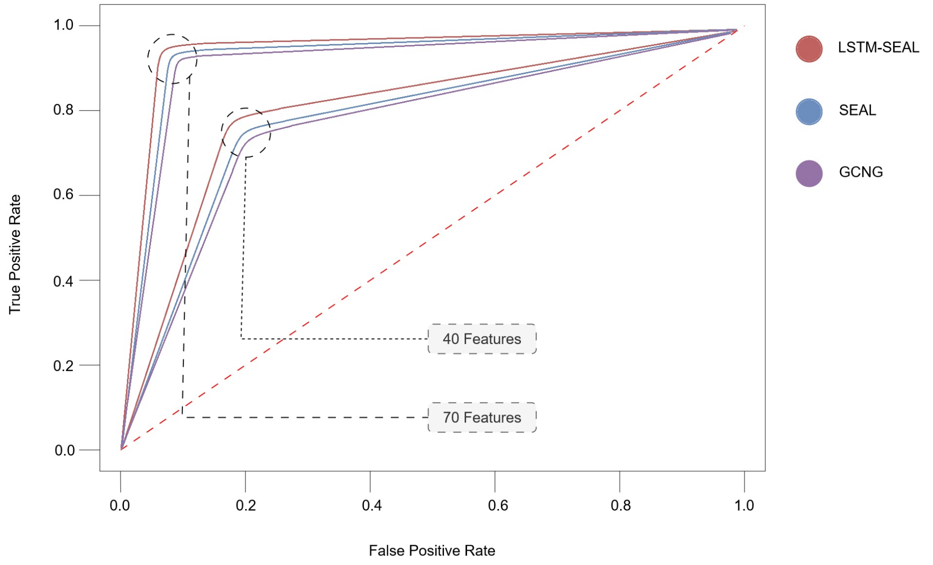


Fig. 2.6.1: ROC graph for SeqFISH+ dataset compared with different approaches for 70 features and 40 features using backward search feature selection method.

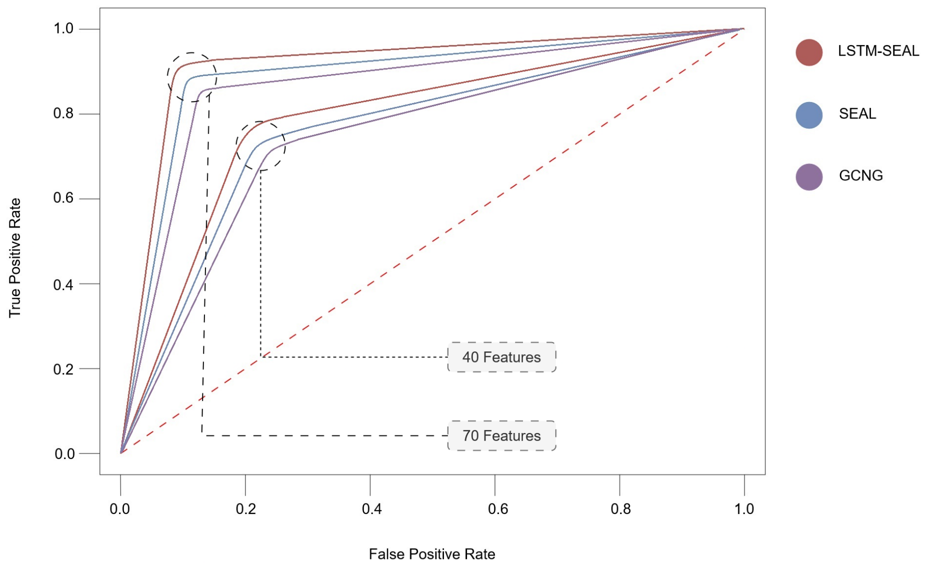


Fig. 2.6.2: ROC graph for MERFISH dataset compared with different approaches for 70 features and 40 features using backward search feature selection method.

The AUROC graphs visually represent the true positive rate against the false positive rate. Figure 2.6.1, illustrates the AUROC graph for SeqFISH+, showcasing the performance of LSTM-SEAL, SEAL, and GCNG at 70 features and 40 features. The graph delineates the nuanced interplay between the true positive and false positive rates, visually understanding the models' discriminative abilities. Similarly, Figure 2.6.2, captures the AUROC graph for the MERFISH dataset, presenting the performance of the three approaches at 70 features and 40 features. This visual representation is a valuable supplement to the numerical AUC values, allowing for a more intuitive grasp of the models' classification accuracy.

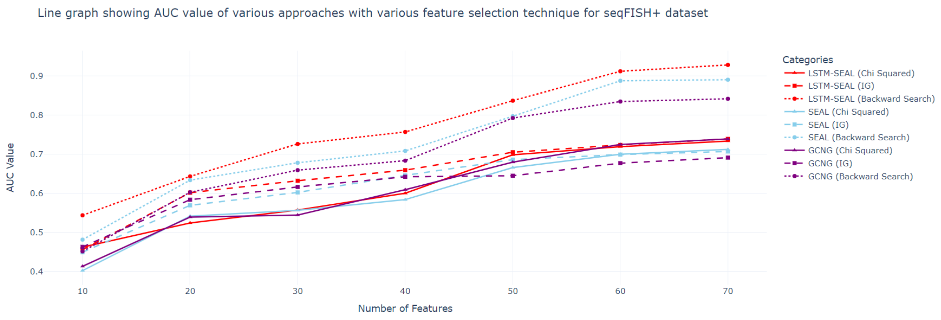


Fig. 2.6.3: Shows a line graph with AUC value of various approaches with different feature selection technique for SeqFISH+ dataset.

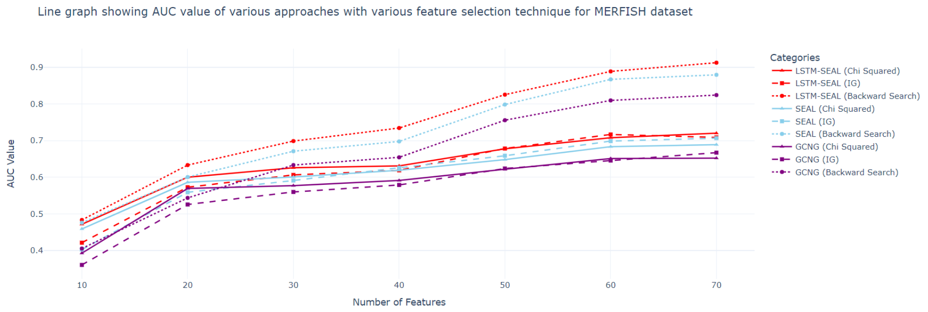


Fig. 2.6.4: Shows a line graph with AUC value of various approaches with different feature selection technique for MERFISH dataset.

Moreover, we have included a line graph to visually represent the quantitative results presented in Table 2.6.1 (Figure 2.6.3). This graphical depiction shows the performance of our approach, LSTM-SEAL, compared to other methods across different feature selection techniques for the seqFISH+ dataset. Similarly, Figure 2.6.4, illustrates the performance trends for the MERFISH dataset presented in Table 2.6.2, allowing for a detailed comparison of our method with other approaches under various feature selection scenarios.

The findings underscore the efficacy of LSTM-SEAL, particularly when coupled with the chi-Squared feature selection method, as it consistently yields superior AUC values. Additionally, the backward search feature selection method demonstrates notable efficacy, especially with LSTM-SEAL, indicating its potential for enhancing model performance.

Our findings demonstrated that the LSTM-SEAL outperforms SEAL and the GCNG. The additional LSTM layer shows a noticeable advantage based on the AUC values compared to SEAL for both datasets. Among the feature selection techniques, backward search outperformed chi-squared and information gain. The combination of LSTM-SEAL with backward search produced some of the best results among the three approaches and feature selection techniques across the two datasets, producing results with AUC values of 0.9281 for the seqFISH+ data and 0.9129 for the MERFISH dataset.

In spatial transcriptomics, crucial for deciphering cell-cell interactions, the SeqFISH+ dataset is notable for evaluating diverse prediction approaches and feature selection techniques. As elucidated by the results in Table 2.6.1, LSTM-SEAL consistently outshines its counterparts, showcasing robust predictive capabilities. Across different feature counts, the AUC scores for LSTM-SEAL demonstrate a steady ascent, culminating at the peak of 70 features. While trailing LSTM-SEAL, SEAL exhibits commendable performance, and GCNG, though competitive, falls short of the superior AUC scores achieved by LSTM-SEAL and SEAL. This underscores LSTM-SEAL's proficiency in integrating spatial information for effective cell-cell interaction prediction.

Shifting the focus to the MERFISH dataset, the dominance of LSTM-SEAL becomes even more pronounced, surpassing other approaches by a significant margin, as evidenced by the results in Table 2.6.2. While competitive, SEAL and GCNG consistently lag behind the above approach. Notably, the efficacy of LSTM-SEAL across both datasets highlights its potential as a robust tool for unraveling intricate cellular communication networks in spatial transcriptomics. This highlights the improvements by including the LSTM layer in the original SEAL.

Delving into the feature selection realm, comparing chi-Squared, information gain, and backward search across approaches and datasets reveals intriguing trends—chi-Squared and information gain exhibit similar patterns, improving performance as the feature count increases. In contrast, backward search is the most compelling feature selection method, consistently delivering superior results, particularly evident in the Merfish dataset, where it achieves the highest AUC scores at 70 features. This accentuates the crucial role of considering feature removal, not just selection, in optimizing predictive models, emphasizing the nuanced nature of feature engineering.

These findings collectively contribute valuable insights to the spatial transcriptomics landscape. LSTM-SEAL's superior performance positions it as a promising candidate for deciphering complex cellular communication networks. The prominence of backward search underscores the significance of thoughtful feature selection strategies, shedding light on the intricate dance between feature selection techniques, prediction approaches, and dataset characteristics. This comprehensive evaluation guides the selection of methodologies for predicting cell-cell interactions. It lays the groundwork for further refinement and exploration in the dynamic and evolving field of spatial transcriptomics and computational biology.

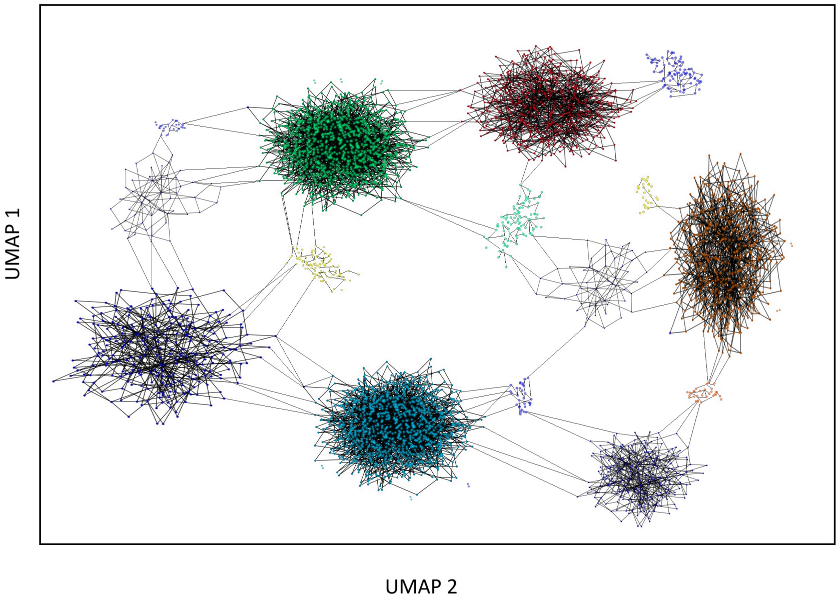


Fig. 2.6.5: Displays a NetworkX visualization of cell-cell interactions in SeqFISH+, using UMAP for spatial awareness. Nodes represent cells, and links signify predicted interactions.

In conjunction with the quantitative evaluation of predictive models, the application of NetworkX, a powerful Python library for network analysis [12], offers a compelling visual dimension to exploring cell-cell interactions. Utilizing UMAP within NetworkX allows for creating a spatially informed network visualization, enhancing our understanding of the cellular communication landscape.UMAP, which stands for Uniform Manifold Approximation and Projection, is a dimensionality reduction technique commonly used in data visualization tasks.UMAP serves the purpose of projecting high-dimensional data into a lower-dimensional space while preserving the local and global structure of the data. The key idea is to represent complex relationships between cells in a more visually interpretable form [15]. The image depicted in Figure 2.6.5, provides an illustrative example derived from the SeqFISH+ dataset. Similarly Figure 2.6.6, illustrates a similar depiction for MERFISH dataset with nodes representing cells, and links signify predicted interactions. This network visualization reveals a complex interplay of cellular interactions, with nodes representing individual cells and links signifying predicted interactions.

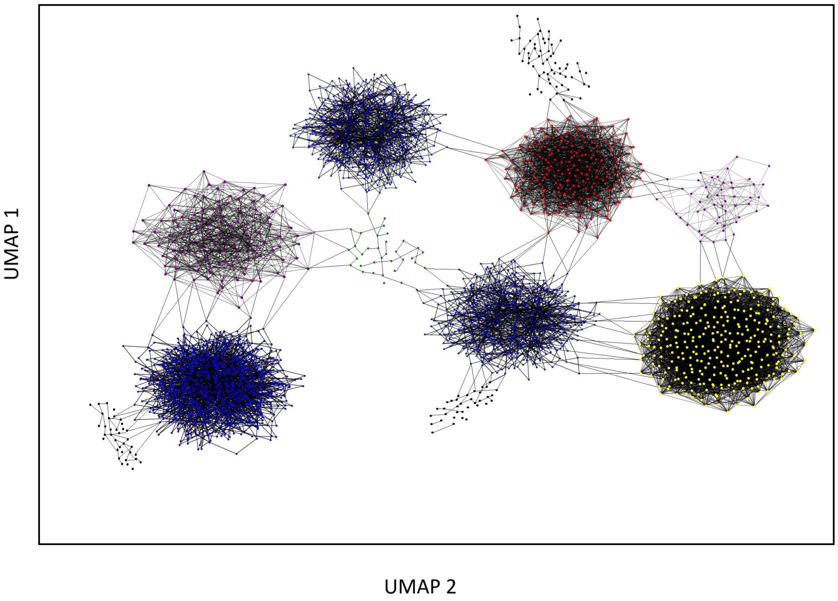


Fig. 2.6.6: Displays a NetworkX visualization of cell-cell interactions in MERFISH dataset, using UMAP for spatial awareness. Nodes represent cells, and links signify predicted interactions.

Upon close examination, it becomes apparent that not all cells are linked, indicating that some cells may have weaker interaction signals. The sparsity of links for specific cells underscores the nuanced nature of cellular behavior in spatial contexts. Furthermore, the network visualization unveils intriguing patterns in the form of dense clusters. These clusters represent groups of cells with strong predicted interactions, suggesting potential biological significance. The formation of dense clusters indicates a higher likelihood of coordinated cellular activities within these groups. On the other hand, less densely connected regions signify areas with fewer or weaker interactions, providing insights into the spatial organization and functional relationships among cells [3].

The integration of NetworkX and UMAP thus serves as a powerful tool for translating predictive model outcomes into visually interpretable representations. This not only aids in identifying critical cellular communication patterns but also facilitates the identification of cellular communities with distinct interaction profiles. Such visualizations provide a holistic perspective, enabling researchers to explore global and local patterns within the spatial transcriptomics dataset.

References

[1] Ahmad Alsahaf et al. “A framework for feature selection through boosting”. In: *Expert Systems with Applications* 187 (2022), p. 115895.

[2] Erick Armingol et al. “Deciphering cell–cell interactions and communication from gene expression”. In: *Nature Reviews Genetics* 22.2 (2021), pp. 71–88.

[3] Mihir Bafna, Hechen Li, and Xiuwei Zhang. “CLARIFY: cell–cell interaction and gene regulatory network refinement from spatially resolved transcriptomics”. In: *Bioinformatics* 39.Supplement_1 (2023), pp. i484–i493.

[4] Zixuan Cang and Qing Nie. “Inferring spatial and signaling relationships between cells from single cell transcriptomic data”. In: *Nature communications* 11.1 (2020), p. 2084.

[5] Zixuan Cang et al. “Screening cell–cell communication in spatial transcriptomics via collective optimal transport”. In: *Nature Methods* 20.2 (2023), pp. 218–228.

[6] Cheng Chen et al. “Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier”. In: *Computers in biology and medicine* 123 (2020), p. 103899.

[7] Yu-Pei Chen et al. “Single-cell transcriptomics reveals regulators underlying immune cell diversity and immune subtypes associated with prognosis in nasopharyngeal carcinoma”. In: *Cell research* 30.11 (2020), pp. 1024–1042.

[8] Xueer Chen et al. “An Instance-Specific Causal Framework for Learning Intercellular Communication Networks that Define Microenvironments Of Individual Tumors”. In: *Available at SSRN 3925258* ().

[9] Changde Cheng et al. “A review of single-cell rna-seq annotation, integration, and cell–cell communication”. In: *Cells* 12.15 (2023), p. 1970.

[10] Roie Cohen et al. “Precise alternating cellular pattern in the inner ear by coordinated hopping intercalations and delaminations”. In: *Science Advances* 9.8 (2023), eadd2157.

[11] Jiarui Feng et al. “Signaling interaction link prediction using deep graph neural networks integrating protein-protein interactions and omics data”. In: *BioRxiv* (2020), pp. 2020–12.

[12] Mehadi Hasan et al. “Protein–Protein Interaction Network Analysis Using NetworkX”. In: *Protein-Protein Interactions: Methods and Protocols*. Springer, 2023, pp. 457–467.

[13] Elham Karimi et al. “Single-cell spatial immune landscapes of primary and metastatic brain tumours”. In: *Nature* 614.7948 (2023), pp. 555–563.

[14] Massimo La Rosa et al. “A Graph Neural Network Approach for the Analysis of siRNA-Target Biological Networks”. In: *International Journal of Molecular Sciences* 23.22 (2022), p. 14211.

[15] Michelle M Li et al. “Contextualizing protein representations using deep learning on protein networks and single-cell data”. In: *bioRxiv* (2023).

[16] Zhuoxuan Li et al. “SpatialDM for rapid identification of spatially co-expressed ligand–receptor and revealing cell–cell communication patterns”. In: *Nature communications* 14.1 (2023), p. 3995.

[17] Lihong Peng et al. “Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies”. In: *Briefings in bioinformatics* 23.4 (2022), bbac234.

[18] Evan S Walsh, H Troy Ghashghaei, and Xinxia Peng. “Feature selection using co-occurrence correlation improves cell clustering and embedding in single cell RNAseq data”. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2021, pp. 751–756.

[19] Juexin Wang et al. “scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses”. In: *Nature communications* 12.1 (2021), p. 1882.

[20] Tao Wang et al. “Single-cell RNA sequencing in orthopedic research”. In: *Bone Research* 11.1 (2023), p. 10.

[21] Aaron J Wilk et al. “Comparative analysis of cell–cell communication at single-cell resolution”. In: *Nature Biotechnology* (2023), pp. 1–14.

[22] Pengyi Yang, Hao Huang, and Chunlei Liu. “Feature selection revisited in the single-cell era”. In: *Genome Biology* 22 (2021), pp. 1–17.

[23] Wenyi Yang et al. “DeepCCI: a deep learning framework for identifying cell–cell interactions from single-cell RNA sequencing data”. In: *Bioinformatics* 39.10 (2023), btad596.

[24] Yongjian Yang et al. “scTenifoldXct: A semi-supervised method for predicting cell-cell interactions and mapping cellular communication graphs”. In: *Cell Systems* 14.4 (2023), pp. 302–311.

[25] Ye Yuan and Ziv Bar-Joseph. "GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data". In: *Genome biology* 21.1 (2020), pp. 1–16.

[26] Rizgar Zebari et al. "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction". In: *Journal of Applied Science and Technology Trends* 1.2 (2020), pp. 56–70.

[27] Yuansong Zeng et al. "Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks". In: *Briefings in Bioinformatics* 23.5 (2022), bbac297.

[28] Muhan Zhang and Yixin Chen. "Link prediction based on graph neural networks". In: *Advances in neural information processing systems* 31 (2018).

CHAPTER 3

Conclusion and Future Work

3.1 Conclusion

In this research, we introduced a new methodlogy for predicting cell-cell communication in spatial trascriptomics data. Our approach combines graph neural nrtwork with Long Short-Term Memory (LSTM) improving upon the performace of existing approaches. The main focus of our methodology is to unravel intricate cell relationships and predict meaningful interactions. Our approach combines LSTM, great for understanding sequences, with graph neural network to tackle complex graph-structured data. We use the combination of LSTM and a graph neural network to better understand how cells interact, improving our model's ability to predict accurately. Using Backward Search, we chose 70 important features to make the model precise and efficient.

Quantitative evaluation of our model showcases its remarkable performance with high Area Under the Curve (AUC) values of 0.92 for seqFISH+ and 0.91 for MERFISH. These metrics highlight the performance of our methodology in navigating the high-dimensional and sparse nature of spatial transcriptomics data. The careful selection of features, incorporating Backward Search to optimize the choice of features, contributes to the model's efficiency by tackling overfitting and enhancing interpretability. We visualize the results using UMAP and NetworkX in Python, helping us see how cells are connected in space. This visual part reveals dense clusters of strong interactions and less crowded areas, giving a detailed look into how cells behave in different spatial contexts.

3.1.1 Contributions

The main contributions of this thesis can be summarized as follows:

- We introduced an innovative pipeline for predicting cell-cell interactions. Our methodology is designed to handle various spatial transcriptomics datasets effectively, which can be challenging due to their high dimensionality and sparsity.
- We strategically combined Long Short-Term Memory (LSTM) with a graph neural network; our method goes beyond conventional approaches, offering a nuanced understanding of cellular relationships. It improved upon the existing methodology based on graph neural networks to produce the best results.
- We applied the proposed method to different spatial transcriptomics datasets. Additionally, we improved the existing methodologies (like GCNG), extending them by using other feature selection techniques to select the best features (genes). We provided a detailed outlook of our methodology by comparing it against the existing methods to highlight the improvement it provides.
- Leveraging the power of NetworkX and UMAP, we visually mapped cellular connectivity within spatial transcriptomics data. These visual insights contribute to a deeper understanding of the cellular landscape by highlighting different cell clusters with strong cell-cell interactions in the cell network.

3.2 Future Work

The following is a summary of some of the tasks that could be taken into consideration and aid in advancing the domain's research:

- We aim to use improved ground truth data, recognizing its pivotal role in refining the model's performance. Moving beyond negative links generated from non-interacting cells, we aim to establish a more accurate ground truth, fortifying the model's capacity to discern true positives and negatives. This refinement addresses uncertainties in predicting cell-cell interactions, contributing to the model's reliability and robustness [1].

- We plan to broaden the scope of our research by exploring additional datasets from various tissues. This expansion aims to provide a more comprehensive evaluation of our adaptable approach across different biological contexts. Testing the model on diverse datasets enhances the generalization and ensures its effectiveness in capturing spatial relationships.

- A crucial aspect of our future work involves a thorough comparative analysis with other existing approaches in the field. By benchmarking our method against alternative models, we seek to validate its efficacy and identify unique strengths. This comparative approach ensures a well-rounded understanding of our method's performance, fostering advancements in cell-cell interaction prediction and related graph-based tasks.

- A future route for extension involves integrating the concepts of heterophily and homophily into our graph models, further enhancing our understanding of diverse cell interactions in spatial transcriptomics datasets.

- We plan to analyze the selected genes (features) to better understand the key genes found that contribute to the cell cell communication for biological validation

References

[1] Ye Yuan and Ziv Bar-Joseph. "GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data". In: *Genome biology* 21.1 (2020), pp. 1–16.

VITA AUCTORIS

NAME:	Karan Kashyap
PLACE OF BIRTH:	Phagwara, Punjab, India
EDUCATION:	Bachelor of Technology (Computer Science and Engineering), Lovely Professional University, Punjab, India, 2020
	M.Sc. Computer Science, University of Windsor, Windsor, Ontario, Canada, 2021