

Master Thesis



Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science and Engineering

Application of graph neural networks on circRNA-disease association prediction

Silvia Goldasová

Supervisor: Petr Ryšavý, Ph.D.
May 2024

I. Personal and study details

Student's name: **Goldasová Silvia** Personal ID number: **492317**
Faculty / Institute: **Faculty of Electrical Engineering**
Department / Institute: **Department of Computer Science**
Study program: **Open Informatics**
Specialisation: **Bioinformatics**

II. Master's thesis details

Master's thesis title in English:

Circular RNA disease association prediction

Master's thesis title in Czech:

Predikce asociací mezi cirkulárními RNA a nemocemi

Guidelines:

Bibliography / sources:

- [1] Yuwei Guo, Ming Yi, THGNCD: circRNA–disease association prediction based on triple heterogeneous graph network, Briefings in Functional Genomics, 2023;, elad042, <https://doi.org/10.1093/bfgp/elad042>
[2] Guanghui Li, Jiawei Luo, Diancheng Wang, Cheng Liang, Qiu Xiao, Pingjian Ding, Hailin Chen, Potential circRNA-disease association prediction using DeepWalk and network consistency projection, Journal of Biomedical Informatics, Volume 112, 2020, 103624, ISSN 1532-0464, <https://doi.org/10.1016/j.jbi.2020.103624>.
[3] Ryšavý, P., Kléma, J. & Merkerová, M.D. circGPA: circRNA functional annotation based on probability-generating functions. BMC Bioinformatics 23, 392 (2022). <https://doi.org/10.1186/s12859-022-04957-8>
[4] Zhou, Jie, et al. "Graph neural networks: A review of methods and applications." AI open 1 (2020): 57-81.

Name and workplace of master's thesis supervisor:

Bc. Petr Ryšavý, MSc., Ph.D. Intelligent Data Analysis FEE

Name and workplace of second master's thesis supervisor or consultant:

Date of master's thesis assignment: **31.01.2024** Deadline for master's thesis submission: **24.05.2024**

Assignment valid until: **21.09.2025**

Bc. Petr Ryšavý, MSc., Ph.D.
Supervisor's signature

Head of department's signature

prof. Mgr. Petr Páta, Ph.D.
Dean's signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

Date of assignment receipt

Student's signature

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Goldasová** Jméno: **Silvia** Osobní číslo: **492317**
Fakulta/ústav: **Fakulta elektrotechnická**
Zadávající katedra/ústav: **Katedra počítačů**
Studijní program: **Otevřená informatika**
Specializace: **Bioinformatika**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Predikce asociací mezi cirkulárními RNA a nemocemi

Název diplomové práce anglicky:

Circular RNA disease association prediction

Pokyny pro vypracování:

Grafové neuronové sítě představují užitečný nástroj pro učení nad problémy, které lze reprezentovat pomocí grafů. Hlavní jejich výhodou oproti klasickému učení nad numerickými vektory je ve schopnosti řešit problémy jako je například doplňování chybějících hran v grafu. V bioinformatice je jedním z takovýchto problémů predikce asociací mezi cirkulárními RNA a nemocemi.

- 1) Proveďte rešerši grafových neuronových sítí.
- 2) Nastudujte problém predikce asociací cirkulárních RNA a nemocí, metody jejich řešení a zmiňte další příbuzné problémy.
- 3) Navrhněte vhodné datové zdroje, které by šlo zpracovat pomocí grafových neuronových sítí.
- 4) Navrhněte přístup založený na grafových neuronových sítích, který bude řešit problém z bodu 2.
- 5) Implementujte návrh z bodu 4.
- 6) Porovnejte přístup z bodu 5. s vhodnými metodami vybranými z bodu 2.

Seznam doporučené literatury:

- [1] Yuwei Guo, Ming Yi, THGNCD: circRNA–disease association prediction based on triple heterogeneous graph network, Briefings in Functional Genomics, 2023;, elad042, <https://doi.org/10.1093/bfpg/elad042>
- [2] Guanghui Li, Jiawei Luo, Diancheng Wang, Cheng Liang, Qiu Xiao, Pingjian Ding, Hailin Chen, Potential circRNA-disease association prediction using DeepWalk and network consistency projection, Journal of Biomedical Informatics, Volume 112, 2020, 103624, ISSN 1532-0464, <https://doi.org/10.1016/j.jbi.2020.103624>.
- [3] Ryšavý, P., Kléma, J. & Merkerová, M.D. circGPA: circRNA functional annotation based on probability-generating functions. BMC Bioinformatics 23, 392 (2022). <https://doi.org/10.1186/s12859-022-04957-8>
- [4] Zhou, Jie, et al. "Graph neural networks: A review of methods and applications." AI open 1 (2020): 57-81.

Jméno a pracoviště vedoucí(ho) diplomové práce:

Bc. Petr Ryšavý, MSc., Ph.D. Intelligent Data Analysis FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **31.01.2024**

Termín odevzdání diplomové práce: **24.05.2024**

Platnost zadání diplomové práce: **21.09.2025**

Bc. Petr Ryšavý, MSc., Ph.D.
podpis vedoucí(ho) práce

podpis vedoucí(ho) ústavu/katedry

prof. Mgr. Petr Páta, Ph.D.
podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomantka bere na vědomí, že je povinna vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

Datum převzetí zadání

Podpis studentky

Acknowledgements

I would like to express my deepest gratitude to my supervisor for the acceptance of the role of my supervisor and provision of such an engaging and very relevant topic of the circular RNAs and the extremely practical graph neural networks. It is the perfect culmination of my studies at CTU, during which I gathered knowledge that has altered my very perception of world and approach to problem solving. This endeavor would not have been possible without my brother, who introduced me to programming, and the study officer for my programme, Ms. Fialová, who always managed to find a solution to my last minute requests. Also, I could not have undertaken this journey without my close friends, who keep me grounded, inspire me to always strive for more and enrich my life in ways I have not thought to be possible.

Declaration

I declare that this thesis is all my own work and I have cited all sources I have used in the bibliography.

Prague, May 24, 2024

Prohlašuji, že jsem předloženou práci vypracovala samostatně, a že jsem uvedla veškerou použitou literaturu.

V Praze, 24. května 2024

Abstract

In recent years, the world of non-coding RNAs has expanded to reveal the significance of a previously little explored class of molecules – circular RNAs (circRNAs). They show potential as biomarkers in disease diagnosis, for example for diabetes, Alzheimer's disease and numerous types of cancer, since they exhibit different expression levels when comparing transcriptome of diseased and healthy subjects, have long half-life and are conserved. However, determining associations between circRNAs and diseases experimentally is a laborious task. For that reason, the computational approach stands out as a suitable alternative.

This work formulates the problem of a circRNA and disease association prediction as a link prediction task on a graph with nodes representing circRNAs, diseases, miRNAs and genes, with the edges illustrating associations and interactions between these entities. A graph neural network model based on a GraphSAGE representation learning framework was designed and implemented to solve the task. The thesis concludes that graph neural networks are, in comparison with the other methods, a suitable approach for prediction of associations between circular RNAs and diseases.

Keywords: graph neural networks, circRNAs, GraphSAGE

Supervisor: Petr Ryšavý, Ph.D.
Karlovo náměstí 13,
120 00 Prague 2,
Czech Republic

Abstrakt

V posledních letech se svět nekódujících RNA rozšířil a odhalil význam dříve málo prozkoumané třídy molekul – cirkulárních RNA (circRNA). Ukazují potenciál jako biomarkery v diagnostice onemocnění, například pro diabetes, Alzheimerovu chorobu a řadu typů rakoviny, protože vykazují různé úrovně exprese při srovnání transkriptomu nemocných a zdravých subjektů, mají dlouhý polčas rozpadu a jsou konzervované. Experimentální stanovení asociací mezi circRNA a nemocemi je však pracný úkol. Z tohoto důvodu se výpočetní přístup jeví jako vhodná alternativa.

Tato práce formuluje problém predikce circRNA a asociace nemocí jako úlohu predikce spojení na grafu s uzly reprezentujícími cirkulární RNA, nemoci, miRNA a geny, přičemž hrany ilustrují asociace a interakce mezi těmito entitami. Pro řešení úlohy byl navržen a implementován grafový model neuronové sítě založený na reprezentačním výukovém rámci GraphSAGE. Práce dochází k závěru, že grafové neuronové sítě jsou ve srovnání s ostatními metodami vhodným přístupem pro predikci asociací mezi cirkulárními RNA a nemocemi.

Klíčová slova: grafové neuronové sítě, cirkulární RNA, GraphSAGE

Překlad názvu: Aplikace grafových neuronových sítí na predikci asociací circRNA-nemoci

Contents

1 Introduction to circRNAs	1	6 Results and Discussion	35
1.1 Motivation	1	6.1 Results	35
1.2 Types of circRNAs	2	6.1.1 General Evaluation	35
1.3 Characteristics of circRNAs	4	6.1.2 Training Splits	36
1.4 Biological Roles of circRNAs	5	6.1.3 Case Study	37
1.5 CircRNA Applications	5	6.2 Discussion	38
1.6 Chapter Summary	6	6.2.1 Evaluation of the Proposed Model	39
2 Graph Neural Networks Theory	7	6.2.2 Limitations of the Model	39
2.1 Link Prediction	8	6.2.3 Future Work	40
2.2 Graph Neural Networks	8	7 Conclusion	41
2.2.1 Graph Neural Network Models/Architectures	10	Bibliography	43
2.3 Graph Neural Networks Overview	13	A GNN Model	53
3 Related Work	15		
3.1 Network propagation-based methods	15		
3.2 Path-based methods	17		
3.3 Machine learning methods	17		
3.3.1 Matrix factorization-based methods	18		
3.3.2 Deep learning-based methods	19		
3.4 Related Problems	20		
3.5 Summary	20		
4 Data Sources	23		
4.1 Introduction	23		
4.2 Databases	23		
4.2.1 CircRNA-disease association Data Sources	23		
4.2.2 CircRNA-miRNA Association Data Sources	24		
4.2.3 MiRNA-disease Association Data Sources	24		
4.2.4 Gene Related Data Sources	24		
4.2.5 Disease Related Data Sources	24		
4.2.6 CircRNA Related Data Sources	25		
4.3 Challenges	25		
5 Method	27		
5.1 CircRNA-Disease Network Design	27		
5.1.1 Data Preparation Challenges	28		
5.1.2 Data Preprocessing	29		
5.2 Graph Neural Network Model	31		
5.2.1 Loss function	32		
5.2.2 Implementation	32		

Figures

1.1 circRNA splicing, taken from [18]	3
1.2 The roles and effects of circRNAs on pancreatic islet β cells, taken from [83]	5
2.1 Architecture of a variational graph autoencoder, taken from [32]	13
5.1 Illustration of the network design, edges are labelled with the data sources for the respective interactions	28
5.2 Node degree distribution of circRNA nodes	31
5.3 Node degree distribution of disease nodes	31
5.4 GNN model design	32

Tables

5.1 A list of databases for various association types	28
5.2 A list of edge types in the network and their counts	30
5.3 A list of node types in the network and their counts	30
6.1 Results from five evaluation metrics: Area Under the Curve (AUC), Area Under the Precision-Recall Curve (AUPRC), Accuracy (ACC), Recall (REC), Precision (PRE)	36
6.2 Results from the training splits evaluation, evaluated with five evaluation metrics: area Under the Curve (AUC), area under the precision-recall curve (AUPRC), accuracy (ACC), recall (REC), precision (PRE)	37
6.3 A ranked list of associations predicted by the model that were marked as false positives based on the input dataset	37

Chapter 1

Introduction to circRNAs

Francis Crick's assertion in 1957 that the primary function of genetic material is to orchestrate protein synthesis through a two-step process: DNA to RNA, then RNA to proteins [12], marked a transformative era in molecular biology. This period was characterized by groundbreaking discoveries of various RNA molecules intricately involved in the synthesis of proteins, including ribosomal RNA (rRNA), messenger RNA (mRNA), and transfer RNA (tRNA). The prevailing understanding at the time was that RNAs were linear molecules terminated with a 5' and 3' end. This notion shaped the trajectory of biomolecular research and experiment design and inadvertently obscured an entire class of RNA molecules: circular RNAs.

Even with the sporadic and serendipitous discoveries of circular RNAs, they remained disregarded as nonspecific byproducts of the normal splicing process [13] with no biological relevance for decades. The breakthrough finally came in 2012 and 2013 with a series of published papers with telltale names: "Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types" by Salzman et al. [31], "Circular RNAs are abundant, conserved, and associated with ALU repeats" by Jeck et al. [59], "Circular RNAs are a large class of animal RNAs with regulatory potency" by Memczak et al. [52]. These studies identified thousands of circular RNAs, hypothesized about their regulatory function and described their various characteristics, such as their stability, relative abundance in comparison with their linear counterparts, and last but not least, conservation across species. These revelations were made possible mainly thanks to the high-throughput RNA sequencing and development of circRNA-specific bioinformatics algorithms.

1.1 Motivation

The interest in the circRNAs spurred by the findings led to the discovery of their biological role. The research showed that through interaction with DNA, RNA and proteins, circRNAs act as regulators of gene expression, more specifically, by affecting transcription and splicing in nucleus [47] as well as translation and signaling pathways in the cytoplasm. Overall, circRNAs play a role in a large variety of physiological and pathological biological

processes from cell proliferation and death through cell metabolism to immune response [48, 81]. Importantly, numerous studies reported that abnormal expression and mutation of circRNAs plays a role in development of various diseases: atherosclerosis [9], cancer [2, 3], neurodegenerative diseases such as Alzheimer's disease and Parkinson's disease [86] and many others. It can be expected that knowing the involvement of a circRNA in disease development, progression and resolution would help us understand the diseases themselves better. Moreover, this relationship makes circRNAs promising candidates for disease biomarkers and therapeutic targets.

The progress in the circRNA research comes to fruition, as thousands of circRNAs have been discovered. However, it is no longer feasible to experimentally search for possible associations between circRNAs and diseases, since the experimental methods are laborious and expensive. One of the very effective alternatives for discovering new associations lies in employing computational approaches to look for possible candidate associations, which can be later validated through laboratory experiments. This work will focus on one of the newer approaches, i.e. graph neural networks.

In the rest of this chapter, various characteristics and functions of circRNAs contributing to the understanding of the circRNA-disease relationship will be outlined.

1.2 Types of circRNAs

- The term circular RNAs encompasses several types of them, which vastly differ in their roles, host organism domains, biogenesis and composition. The following circRNA classification is adopted from the review on circular RNAs by Lasda et al. [40]:

1. Circular RNA genomes

Genomes of viroids, which are infectious plant pathogens, and hepatitis delta virus have a form of a circular single-stranded RNA. The loop structure allows for rolling circle RNA replication, in which multiple genomic copies are produced from a single initiation event.

2. Circular RNA introns

Introns are noncoding sections of an RNA transcript, as opposed to exons, which are the sections of RNA transcripts translated into a protein. After transcription, introns are generally spliced out of the transcripts in the form of lariats (circles with a tail).

There are multiple classes of circular introns formed by various processes, namely group I introns, group II introns, circular intronic RNAs and excised tRNA introns.

Both *group I introns* and *Group II introns* are self-splicing ribozymes, meaning that they catalyze their own excision from RNA precursors. Furthermore, some members from either of the groups are genetic mobility elements: while some introns from group I encode homing endonuclease

that catalyzes intron mobility even between organisms, an intron from group II can be incorporated back into DNA or RNA by reverse splicing into a new location. Lastly, circular RNA introns can be found in bacteria, some archaea, some viruses and some eukaryotes.

Circular intronic RNAs (ciRNAs) are produced by eukaryotic splicing, which is catalyzed by the spliceosome. For the most part, the lariat intron created during the splicing is degraded shortly thereafter. However, certain consensus RNA motifs near the 5' splice site and branchpoint can promote conformations that limit debranching. This results in a stable ciRNA molecule composed of only the loop part of the lariat, as the tail is broken down [87]. Such ciRNAs tend to accumulate in the nucleus and influence the expression of their parent genes as well as regulate RNA Pol II transcription [87].

Excised tRNA introns The final intron tRNA processing byproduct in some archaea.

3. Circular RNA processing intermediate The intermediates in rRNA and tRNA processing reaction in archaea also take on circular form. The circularization provides means for rearrangement of the RNA sequence order by circularization of a linear RNA followed by relinearization at a different position.
4. Circular noncoding RNA Non-coding RNAs found in archaea [66, 15].
5. Circular RNA spliced exons produced by backsplicing Backsplicing is a non-canonical splicing process, in which in a downstream splice donor of a pre-mRNA is joined to an upstream splice acceptor, creating a closed loop. The products of backsplicing can be seen on the image below.

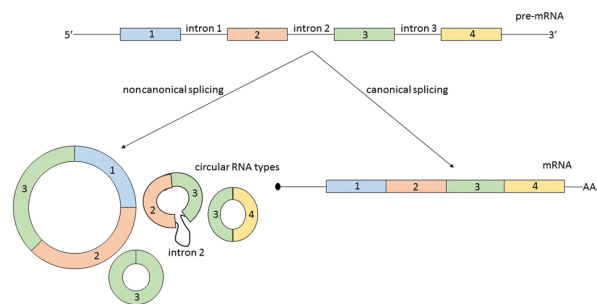


Figure 1.1: circRNA splicing, taken from [18]

Although there is a wide range of RNA molecules with a circular structure, as illustrated by the exhaustive enumeration above, the term *circRNAs* is generally used to only refer to the spliced exons of coding genes produced by backsplicing, as described in the 5th point [4]. The reason for this is that they hold the the most relevance and significance for the human biology and are the most researched as well. This work will also rely on this convention unless specified otherwise.

1.3 Characteristics of circRNAs

Many of the characteristics and behaviour patterns of circRNAs derive from their structure, localization within the cell, biogenesis or length.

Thanks to their circular structure and the ensuing lack of the usual terminal structures: the 5' cap and the 3' polyadenylated tail, circRNAs are intrinsically resistant to RNA decay by exonucleases, which are enzymes that cleave nucleotides one by one from either the 3' or the 5' end of a polynucleotide chain. Presumably, this resistance makes circRNAs stable, which is also confirmed by the experiment by Enuca et al. [19], in which they examined half-lives of 60 circRNAs and their linear counterparts in mammary cells. The experiment revealed that the median half-life of investigated circRNAs is at least 2.5 times longer than the median half-life of their linear counterparts.

CircRNA exhibit tissue specificities and disease specificities. An important characteristic of circRNAs is tissue specificity. Rybak-Wolf et al. analyzed 29 types or stages of neural cells and tissues [57]. Their study revealed that different brain compartments produce different sets of highly expressed circRNAs and that neuronal differentiation is associated with higher circRNA expression.

CircRNA abundance. The tissue specificity impacts also the expression levels of circRNAs. CircRNAs are, for the most part, said to be expressed at levels that are 2–10% of those of their linear counterparts [58, 64], and that they can in some cases surpass the quantities of their linear counterparts even 10-fold [64]. However, such generalization does little for illustration of how much do the expression levels of different circRNAs actually vary. For instance, study [46] states that the ratio of circRNA level to linear RNA level in exosomes was 6-fold higher than that in cells. Study [53] concludes that the ratio of circRNAs with at least as high expression as that of their linear isoforms is 33.66% for blood samples, compared to 18.45% in cerebellum samples and only 9.71% in liver samples.

Conservation. Studies show that mammalian circRNAs tend to be conserved across species. Studying retinal tissues, Chen et al. found that highly expressed mouse circRNAs are more likely to be conserved between mice and humans [8]. Furthermore, Xia et al. found multitude of circRNAs conserved between mouse and fetal human brain, and in small numbers, some were conserved between adult human heart, liver, skin and lung [78].

Location. While intron-containing circRNAs stay in the nucleus, most circRNAs are after their biogenesis transported from the nucleus to the cytoplasm. CircRNAs can be also found in blood [53], saliva [49], plasma [90] or exosomes [46].

Length. The length of circRNAs ranges from approximately 100 to 4000 base pairs long [63]. In the study of circRNA in exosomes, the median length

for the 1215 identified circRNAs was 350 nt. Usually, circRNAs contain 1-5 exons [52].

1.4 Biological Roles of circRNAs

The main function of circRNAs is regulation of post-transcriptional activity by acting as microRNA sponges: microRNAs (miRNAs) bind to the complementary miRNA binding sites on circRNAs, what prevents them from binding to their intended mRNA targets. In this manner, circRNAs can regulate gene expression. This function is supported by the fact that circular RNAs are depleted of polymorphisms at these microRNA binding sites [67]. The mechanism can be illustrated by the description of the circZNF566/miR-4738-3p/TDO2 relationship by the Li et al. [45]: overexpression of the enzyme TDO2 promotes the mobility, migration, invasion, and proliferation of hepatocellular carcinoma cells, miRNA miR-4738-3p acts as a tumor suppressor by directly suppressing TDO2 expression through binding to the 3' untranslated region of TDO2 mRNA, and circZNF566 acts as a miR-4738-3p sponge, attenuating the inhibitory effect of the miRNAs on the expression of TDO2 through this competitive binding.

Furthermore, the circRNAs were shown to also bind RNA-binding proteins, cell cycle regulatory proteins, and participate in the protein complex assembly, what enables them to regulate gene splicing, transcription, translation, or epigenetic regulation. Exceptionally, they even encode small peptides: 46 circRNAs from 37 genes were found to have their corresponding proteins expressed according to mass spectrometry [7].

The functions of circRNAs can be well illustrated by the following figure, which depicts roles of circRNAs relating to the pancreatic islet β cells:

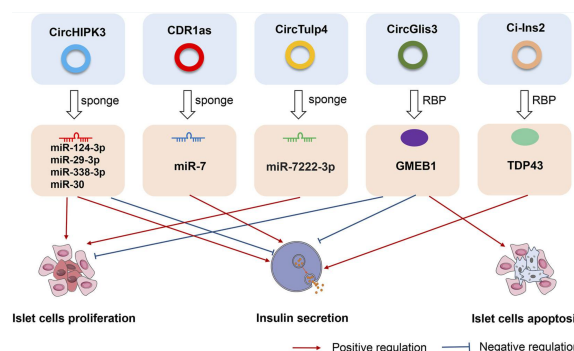


Figure 1.2: The roles and effects of circRNAs on pancreatic islet β cells, taken from [83]

1.5 CircRNA Applications

CircRNAs show great potential as disease biomarkers and novel therapeutic targets as they have tissue-specific expression, long half-life, are conserved,

come in high abundances, are released from cells to the circulatory system, and are associated with all types of diseases: cardiovascular, neurological, metabolic, or immune diseases [65].

For example, a study [34] investigated the relationship between circRNA CDR1as and bile duct cancer by comparing the circRNA expression in 54 paired tumor and adjacent normal tissues of cholangiocarcinoma patients. The data showed that the overexpression of Cdr1as is associated with the advanced cancer stage, lymph node invasion, and postoperative recurrence and lower survival chance. They conclude that cRNA Cdr1as was significantly upregulated in tumor tissues and suggest it as an independent prognostic biomarker for patients with bile duct cancer.

CDR1as is not specific only to the bile duct cancer, as it was found to serve as a mediator in alteration of the tumor microenvironment through its regulatory role of the TGF- β signaling pathway and ECM-receptor interaction [91]. Studies suggest it as a biomarker also for colorectal cancer [73], laryngeal squamous cell carcinoma [85] or triple-negative breast cancer [61].

The diagnostic landscape of certain diseases poses formidable challenges even today, in some cases falling short in providing specific diagnostic tests for a disease. For instance, diagnosing chronic fatigue syndrome is based on the symptoms and on ruling out other conditions that could be causing those symptoms. In this context, the possibility of identification of circRNAs from peripheral blood of chronic fatigue syndrome patients, as shown in [11], proves to be extremely advantageous, since it can finally give us means for more precise diagnostics. Likewise, circRNAs can be used for diagnosing endometriosis, a chronic gynaecological condition also called ‘missed disease’ due to its unclear aetiology and multitude of shared symptoms with other diagnoses [30]. A study [79] found a candidate circ_0002198 with the AUC for distinguishing ovarian endometriosis at 0.846.

■ 1.6 Chapter Summary

To summarize, the circRNAs are naturally occurring ribonucleic acids in a form of a covalently closed loop. Their dysregulation plays a role in multitude of diseases, mostly through the mechanism of miRNA sponging, when by binding to the circRNAs are the miRNAs prevented from binding to their intended mRNA targets.



Chapter 2

Graph Neural Networks Theory

Data encountered in real-world scenarios frequently exhibit a graph-like structure. To investigate graphs and capture the information encoded in the graph structure, various graph analysis techniques have been developed. They can be used to address problems such as predicting potential interactions between proteins, new drug discovery or analysis of biological pathways. Graph analysis is closely connected to the concept of graph representation learning, whose goal is to accurately learn to represent graph nodes, edges or subgraphs by low-dimensional vectors. The specific problems in graph analysis fall usually under one of the four tasks defined on graphs: node classification, link prediction, graph regression, node clustering, or network classification, with graph neural networks being nowadays a popular approach to solving such tasks. Some of the current popular GNN architectures are recurrent graph neural networks, convolutional graph neural networks, graph autoencoders, and spatial-temporal graph neural networks.

To effectively handle and learn from graph-structured data, graph neural networks (GNNs) have emerged as a very successful subgroup of machine learning models. They were motivated by convolutional neural networks (CNNs) and graph embedding methods, both of which had numerous limitations regarding their application in graphs. CNNs share many characteristics imperative also for graph analysis methods, namely, they do a good job at extracting local spatial features, use shared weights and multiple layers. However, they are appropriate only for special instances of graphs, such as images, which can be thought of as fixed-size grid graphs, and are not generalizable to more complex graphs. The disadvantage of graph embedding methods lay in their inability to generalize to unseen graphs and absence of parameter sharing between nodes, meaning that number of parameters increases proportionally to the number of nodes.

This chapter explores the characteristics and essential concepts of graph neural networks, with focus on their application and advancements related to link prediction tasks due to the nature of this work.

inductive bias. Inductive bias of a learning algorithm is the set of assumptions that can guide the model to a better predictive performance, fewer parameters and better generalization [60]. Regarding the graph analysis, the

characteristics that can be leveraged are the relationships across all graph components, t.i. edges, nodes, global, preservation of the explicit relationships - the edges themselves or preservation of graph symmetries (permutation invariance). Specifically, the graph symmetry should be reflected in the design of the utilized transformation, so that the order of operations on nodes or edges would not matter. Therefore, a good graph neural network model has a relational inductive bias.

2.1 Link Prediction

Link Prediction is a task in network analysis, the goal of which is to predict missing or unobserved links between two nodes in the network given structural and feature information. Given a graph $G(V, E)$, where V represents the set of nodes as entities and E represents the set of edges as relationships between the entities, the predictions are made based on integration of information from the network structure, node attributes and network’s existing relationships.

Arrar et al. classify link prediction methods in their comprehensive survey [1] into four categories, namely similarity-based methods, dimensionality reduction-based methods, machine learning technique-based methods and other methods. Since this work explores the application of graph neural networks for the circRNA-disease association prediction, this work will examine only GNNs, which are a subcategory of machine learning technique-based methods.

2.2 Graph Neural Networks

Graph Neural Networks (GNNs) are a class of deep learning methods specifically designed for studying graph-structured data. Sanchez-Lengeling, et al. [60] defines a GNN as an optimizable transformation on all attributes of the graph (nodes, edges, global-context) that preserves graph symmetries (permutation invariances).

An important concept in graph theory is node embedding, which is a projection of nodes from the original network to a low-dimensional embedding space in such a way that nodes that are similar in the original network share embeddings that are proximal in the embedding space.

One key desideratum in the graph neural network design is permutation invariance with respect to the node ordering, meaning that the output of the model remains the same when changing the order in which nodes and edges are processed. This requirement reflects the characteristic of graphs, that, in general, nodes have no default ordering, and therefore the order of operations on nodes or edges should not matter. Furthermore, given that graphs are very flexible structures, graph neural networks also have to address the issue that nodes have a varying number of neighbours and the graphs themselves

can be of vastly different sizes. These challenges are addressed by the concept of neural message passing.

Message Passing Neural Network Framework. Neural Message Passing provides a framework for spatial graph convolution utilized for modelling the complex relationships and dependencies in graph-structured data by propagating information across the graph through a series of message passing iterations. At each iteration, information is aggregated from neighboring nodes and combined with the node’s own features to update its representation. The message passing iteration consists of 3 steps:

- gathering of all the neighboring node embeddings,
- aggregation of all messages from neighboring nodes via an aggregate function,
- update of the node embedding according to the information aggregated from the node’s neighbourhood

It can be formularized as follows:

$$\mathbf{h}_u^{(k+1)} = \gamma^{(k)} \left(\mathbf{h}_u^{(k)}, \bigoplus_{v \in \mathcal{N}(u)} \phi^{(k)} \left(\mathbf{h}_u^{(k)}, \mathbf{h}_v^{(k)}, e_{vu} \right) \right),$$

where $\mathbf{h}_u^{(k)}$ represents a hidden embedding of a node u , the node representation is updated based on information gathered from its neighborhood $\mathcal{N}(u)$. $\gamma^{(k)}$ and $\phi^{(k)}$ denote arbitrary differentiable functions such as multilayer perceptrons, with $\gamma^{(k)}$ representing the update function. \bigoplus represents the aggregation function that is differentiable and permutation invariant, such as sum, mean or max. e_{vn} denotes an optional feature vector of the edge from node v to u .

Node embeddings at iteration $k=0$ are initialized to the input features X_u for all nodes u . After K iterations of message passing, each node’s updated representation contains information of all neighbors up to K -distance, which can be thought of as a subgraph representation. We denote these learned embeddings z_u for each node u as:

$$\mathbf{z}_u = \mathbf{h}_u^{(K)}, \forall u \in \mathcal{V}.$$

By this iterative message passing from nodes to their neighbours, GNNs can encode the local neighborhood information of each node into its learned representation. As a result, the representation of each node will be covering both the structural information of its neighbourhood and the features of the node. It’s important to note that the usage of the appropriate aggregation function guarantees permutation invariance.

A Basic Graph Neural Network. Utilization of the GNN framework can be illustrated by the example of a basic graph neural network as defined in [27] by Hamilton, which itself is a simplification of a GNN model proposed by Scarselli et al.:

$$\mathbf{h}_u^{(k)} = \sigma \left(\mathbf{W}_{self}^{(k)} \mathbf{h}_u^{(k-1)} + \mathbf{W}_{neigh}^{(k)} \sum_{v \in \mathcal{N}(u)} \mathbf{h}_v^{(k-1)} + \mathbf{b}^{(k)} \right),$$

where $\mathbf{W}_{self}^{(k)}, \mathbf{W}_{neigh}^{(k)} \in \mathbb{R}^{d^{(k)} \times d^{(k-1)}}$ represent trainable parameter matrices, $\mathbf{b} \in \mathbb{R}^{d^{(k)}}$ a bias term. The parameters can either be shared across the iterations or trained separately. σ denotes an element-wise non-linearity, e.g. a hyperbolic tangent function (tanh) or rectified linear unit (ReLU) activation function. The message passing in this basic GNN resembles a standard multi-layer perceptron (MLP) in the sense that they both rely on linear operations followed by a single element-wise non-linearity.

Alternatively to the node perspective, GNNs can be also defined by graph-level equations:

$$\mathbf{H}^t = \sigma \left(\mathbf{A} \mathbf{H}^{(t-1)} \mathbf{W}_{neigh}^{(t-1)} + \mathbf{H}^{(t-1)} \mathbf{W}_{self}^{(t)} \right),$$

where $\mathbf{H}^t \in \mathbb{R}^{|V| \times d}$ denotes the matrix of node embeddings in the GNN at layer t , \mathbf{A} is the graph adjacency matrix. The graph-level definition shows that a GNN can be implemented using just a few sparse matrix operations. Regarding the implementation, it is important to also note that it is a common practise to add self-loops to the input graph. As a result, the aggregation function combines the messages from the neighbouring node as well as from the node itself, simplifying the operation, which in turn leads to a decreased chance of overfitting, but also decreased GNN expressivity due to the inability to differentiate between the information coming from the node's neighbours and the node itself [27].

2.2.1 Graph Neural Network Models/Architectures

Graph Convolutional Network. Graph Convolutional Network GCNs utilize the Kipf normalized aggregation and self-loops. The message passing function in the GCN model is defined as:

$$\mathbf{h}_v^{(k)} = \sigma \left(\mathbf{W}^{(k)} \sum_{v \in \mathcal{N}(u) \cup \{u\}} \frac{\mathbf{h}_v^{(k-1)}}{\sqrt{|\mathcal{N}(u)| |\mathcal{N}(v)|}} \right).$$

The normalization of the features of neighbouring nodes is necessary due to the fact that nodes can have large differences in the number of neighbours which could lead to high differences between the feature representations of nodes, which can in turn cause numerical instabilities and difficulties for optimization. Normalization by the degree of the nodes reflects the idea that high-degree nodes are less useful for inferring information as they interact

with too many other nodes, while the low-degree nodes are assumed to have more meaningful interactions for extracting patterns.

Alternatively, the GCN can be described on the graph level as follows [37]:

$$\mathbf{H}^{t+1} = \sigma \left(\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{H}^{(t)} \mathbf{W}^{(t)} \right),$$

where

$$\begin{aligned} \tilde{\mathbf{A}} &= \mathbf{A} + \mathbf{I}_N \\ \tilde{\mathbf{D}}_{uu} &= \sum_v \tilde{\mathbf{A}}_{uv}. \end{aligned}$$

\mathbf{A} denotes the adjacency matrix, which is modified by the addition of a self-connections for each node, resulting in the $\tilde{\mathbf{A}}$ matrix. $\tilde{\mathbf{D}}$ is the diagonal node degree matrix of $\tilde{\mathbf{A}}$, $\mathbf{W}^{(t)}$ represents a layer-specific trainable weight matrix, $\mathbf{H}^{(t)}$ is a matrix holding feature representations from the t^{th} layer, $\mathbf{H}^{(0)}$ is set to the input feature vector \mathbf{X} , $\sigma(\cdot)$ denotes an activation function, such as the $ReLU(\cdot)$. This layer-wise propagation rule for neural network models, presented by Kipf and Welling in 2016, was derived from a first-order approximation of a localized filter of a spectral graph convolution.

Graph Attention Network. Graph Attention Networks (GATs) are a popular variant of GNNs that leverage attention mechanisms in the aggregation step for better feature learning on graphs. The key idea is to assign an attention weight to each neighbor, which weighs the neighbor’s importance in the aggregation of information from neighbouring nodes. The first GAT model, as presented in 2018 by Veličković et al.[68] defines the attention coefficients as follows:

$$e_{uv} = a(\mathbf{W}\mathbf{h}_u, \mathbf{W}\mathbf{h}_v),$$

where e_{uv} , the attention coefficient expressing the importance of the features of node v to the node u , is computed by a self-attention mechanism $a : \mathbb{R}^{F'} \times \mathbb{R}^{F'} \rightarrow \mathbb{R}$. The coefficients are normalized by the softmax function across all choices of node v , so that they can be used for comparisons across different nodes:

$$\alpha_{uv} = \text{softmax}_v(e_{uv}) = \frac{\exp(e_{uv})}{\sum_{k \in \mathcal{N}(u)} \exp(e_{uk})}.$$

In practice, the paper uses a single-layer feedforward neural network as their attention mechanism a parametrized by a weight vector $\mathbf{a} \in \mathbb{R}^{2F'}$, treated with leakyReLU nonlinearity:

$$\alpha_{uv} = \frac{\exp(\text{leakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_u \parallel \mathbf{W}\mathbf{h}_v]))}{\sum_{k \in \mathcal{N}(u)} \exp(\text{leakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_u \parallel \mathbf{W}\mathbf{h}_k]))},$$

where \parallel denotes the concatenation operation. The aggregation itself is computed as a linear combination of features of neighbouring nodes and their

corresponding normalized attention coefficients:

$$\mathbf{h}_u^{(k)} = \sigma \left(\sum_{v \in \mathcal{N}(u)} \alpha_{uv} \mathbf{W} \mathbf{h}_v^{(k-1)} \right).$$

To improve the model’s performance and stabilize the learning process of self-attention, the attention mechanism can be extended to multi-head attention, which employs K independent attention mechanisms to calculate the aggregation. These K transformations are then concatenated or averaged to obtain a new node representation. To enumerate some of the graph attention networks properties, they are computationally efficient as the computation of attention coefficients can be parallelised, localised, storage efficient, have a fixed number of parameters and can implicitly specify neighbour importances.

Graph Sample and Aggregate (GraphSAGE). GraphSAGE (SAmple and aggreGatE) [28] is a general inductive framework for node embedding based on learning how to aggregate feature information from a node’s local neighborhood. The key idea of the approach is that instead of training individual embeddings for each node, a function that generates embeddings by sampling and aggregating features from a node’s local neighborhood is trained. Due to this design, GraphSAGE can predict embeddings of unseen nodes without needing re-training. Hamilton et al. [28] define the message passing in the GraphSAGE framework as follows:

$$\begin{aligned} \mathbf{m}_{\mathcal{N}(u)}^{(k)} &= \text{AGGREGATE}_k \left(\left\{ \mathbf{h}_v^{(k-1)}, \forall v \in \mathcal{N}(u) \right\} \right) \\ \mathbf{h}_u^{(k)} &= \sigma \left(\mathbf{W} \cdot \text{CONCAT}(\mathbf{h}_u^{(k-1)}, \mathbf{m}_{\mathcal{N}(u)}^{(k)}) \right), \end{aligned}$$

where firstly, the messages from neighbouring nodes are aggregated and afterwards concatenated with the node’s embedding, what keeps the two information separate rather than mixing them up. The paper examined three candidate aggregator functions, namely *mean aggregator*, which is a rough, linear approximation of a localized spectral convolution [28]; *LSTM aggregator*, which has a larger expressive capacity, but since LSTMs are not inherently symmetric, the framework applies the LSTMs to a random permutation of the node’s neighbours in order to emulate the permutation invariance; and lastly the *pooling aggregator*, which uses a fully-connected neural network to firstly make a transformation of the embeddings of the neighbouring nodes and then applies an elementwise max-pooling operation.

Graph Autoencoder. Variational Graph Autoencoder (VGAE) [35] The main idea of a variational graph autoencoder [36] is that it embeds the input \mathbf{X} to a distribution rather than an embedding vector. From this distribution, a random sample \mathbf{z} can be sampled.

The encoder (inference model) of VGAE consists of two GCN layers. The inputs consists of an adjacency matrix A and a feature matrix X . The first

GCN layer generates a lower-dimensional feature matrix. The second GCN layer generates μ and $\log\sigma^2$, where z can be calculated as $z = \mu + \sigma * \epsilon$, where $\epsilon \sim N(0, 1)$. The decoder (generative model) uses an inner product between latent variables. The decoder output is a reconstructed adjacency matrix $\hat{A} = \sigma(zz^T)$, where σ is a logistic sigmoid function.

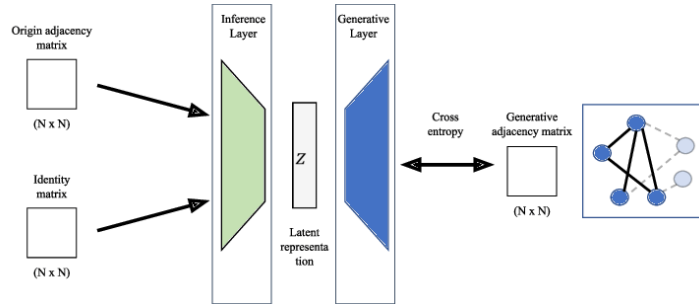


Figure 2.1: Architecture of a variational graph autoencoder, taken from [32]

2.3 Graph Neural Networks Overview

Graph neural networks perform well at capturing complex patterns encoded in the graph structure and can therefore leverage both the structural and functional information of the samples. However, they face computational challenges when applied to large-scale networks, can be constrained by the limited memory resources, and may exhibit biases when dealing with imbalanced node or edge distributions.

Chapter 3

Related Work

This chapter aims to provide an overview of the computational methods employed in the deciphering the complex relationship between circRNAs and various diseases.

While the principles of the methods wildly differ, many of them share common characteristics, such as they work with heterogeneous graphs. Heterogeneous graphs can have nodes and edges of different types, what makes them better suited for representation of real-world data. The methods also often use same techniques for feature extraction and the data sources also do not differ much. However, the approaches themselves are vastly different, and can be classified into 3 larger groups: network propagation-based methods, path-based methods, and machine learning methods.

3.1 Network propagation-based methods

Biological networks frequently serve as models for depicting relationships among biological entities. To solve biological network problems, various network propagation-based methods have been developed. The principle of the methods is to iteratively diffuse input node data along edges so that nodes accumulate also information about the structure of the network in their neighbourhood. This category includes methods based on random walks and label propagation.

Network propagation can be also viewed as a special case of graph convolutions [69], which is a mathematical operation frequently employed in graph neural networks.

BRWSP algorithm: Biased Random Walk. BRWSP algorithm [44] constructs a heterogeneous network HN by integrating known circRNA-disease associations, circRNA coexpression similarity network, gene functional similarity network, disease similarity network, circRNA-gene associations, and gene-disease associations.

The algorithm uses a biased random walk to find paths in the network between a certain circRNA and a disease. The strategy of selecting the next node in the walk is controlled by the parameter q , by which we can either set

a preference for selecting nodes within the same community or to exploring in order to get a macro view of the neighbourhood. The strategy itself is formalized as follows:

$$P(c_{k+1} = x | c_k = v, c_{k-1} = t) = \begin{cases} \frac{\Theta(t,v,x) * HN(v,x)}{\sum_{i \in Nei(v)} HN(v,i)}, & x \in Nei(v), x \notin path, \\ 0, & otherwise, \end{cases}$$

$$\Theta(t, v, x) = \begin{cases} q, & x \in Nei(v) \text{ and } x \in Nei(t), \\ 1 - q, & otherwise, \end{cases}$$

where $P(c_{k+1} = x | c_k = v, c_{k-1} = t)$ represents the transition probability of selecting node x as the next node in the biased random walk, with node v being the currently visited node and node t being the last visited node. $Nei(v)$ and $Nei(t)$ represent the neighbourhoods of nodes v and t , respectively.

The predicted association score for circRNA $c(i)$ and disease $d(j)$ is then calculated as:

$$score(c(i), d(j)) = \sum_{path_i \in all_paths} \left(\prod_{e=1}^{len(path_i)} W_e(path_i) \right)^{\alpha \cdot len(path_i)}$$

where $score(c(i), d(j))$ represents the score for the circRNA-disease pair association, $all_paths = \{path_1, path_2, \dots, path_n\}$ represents the searched paths between circRNA $c(i)$ and disease $d(j)$, $path_i$ represents the i th searched path, $W_e(path_i)$ represents the weight of the e th edge in the $path_i$, and lastly, the parameter α represents a decay factor. The function is designed in a way that it assigns higher scores to circRNA-disease pairs connected by numerous paths of shorter length consisting of edges of higher weights.

RWRKNN algorithm: Random Walk with Restart and k-Nearest Neighbor.

Inspired by the application of the random walk with restart (RWR) and k-nearest-neighbor (kNN) algorithm for identification of drug-target interactions [41], Xiujuan Lei and Chen Bian applied a similar approach for identification of novel circRNA-disease associations [42].

The method consists of four steps: construction of association and similarity matrices for circRNAs and diseases separately, RWR for each circRNA and disease, feature weighting and kNN model training.

The random walk with restart performed on association matrices estimates an affinity score between a circRNA (disease) seed node and all other circRNA (disease) nodes. The algorithm starts in the seed node and in each iteration, a random walker either explores a neighbour node with probability $1 - c$ or restarts from the seed node with probability c :

$$\mathbf{p} = (1 - c)\mathbf{W}\mathbf{p} + c\mathbf{q},$$

where \mathbf{W} is a normalized adjacency matrix, \mathbf{q} is the starting vector, and \mathbf{p} represents the steady-state probabilities reached after some iterations.

Generally, nodes interconnected with the seed node obtain higher scores than nodes that are further away in the network. The point of restarting is to prevent the local accumulation of resources in distant subnetworks. The affinity scores, expressing network topology information, are used to weight circRNA and disease features.

The kNN model is trained with positive and negative samples of the weighted features of circRNA-disease pairs and uses Minkowski distance metric. The algorithm predicts the associations of the sample by looking at the classes of the k nearest neighbors.

3.2 Path-based methods

Path-based methods are concerned with the characteristics of the possible paths between a certain circRNA-disease pair in a constructed graph. More specifically, they calculate the association score between the circRNA and the disease from number of paths between the nodes and the path lengths. [21].

KATZHCA algorithm. KATZHCA method [21] construct a heterogeneous network by integrating known circRNA-disease associations, circRNA expression profile similarity, disease phenotype similarity and Gaussian interaction profile kernel similarity. The adjacency matrix for the network has a form

$$\mathbf{A}^* = \begin{bmatrix} \mathbf{SC} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{SD} \end{bmatrix},$$

where \mathbf{SC} is the circRNA similarity network, \mathbf{SD} is the disease similarity network, and \mathbf{A} represents the adjacency matrix of known circRNA-disease associations.

In this network, the score for a circRNA-disease pair is calculated as integration of the all the walks of different lengths between the circRNA and disease nodes:

$$S(c(i), d(j)) = \sum_{l=1}^k \gamma^l \mathbf{A}^{*l}(i, j)$$

The parameter γ is used to account for the decrease of the significance of longer walks.

3.3 Machine learning methods

The category of machine learning-based approaches employed in predicting novel circRNA-disease associations is diverse, encompassing methods such as DWNN-RLS [80], which uses regularized least squares of Kronecker product kernel, RWLR [17] method makes predictions through binary logistic regression model, and GBOTCDA method [43], which uses gradient boosted decision trees. This category also contains two larger groups of methods: matrix factorization-based and deep learning-based methods.

3.3.1 Matrix factorization-based methods

Matrix factorization-based methods can be used for solving a task of filling in the missing values of a partially observed matrix. The principle of them is to find two low-dimensional matrices such that their factorization will approximate the original, partially observed, input matrix.

iCircDA-MF algorithm. iCircDA-MF method [72] begins with the construction of the similarity matrix of diseases from disease semantic information and known circRNA-disease associations, and construction of the similarity matrix of circRNAs from known associations of types circRNA-gene, gene-disease and circRNA-disease. In order to take into account not only the known circRNA-disease associations, our knowledge of which is still limited, but also consider the likely associations, an alternative circRNA-disease association adjacency matrix is constructed based on the circRNA and disease similarity matrices. This step is based on the assumption that similar circRNAs will be associated with similar diseases. The expanded circRNA-disease association adjacency matrix \mathbf{A}' combines interaction profiles for circRNAs and diseases, where the interaction profiles of circRNAs (diseases) are calculated as a normalized weighted sum of the interaction profiles of the k most similar circRNAs (diseases) in the original association matrix.

iCircDA-MF algorithm performs the basic non-negative matrix factorization to predict the non-negative association scores between circRNA-disease pairs. Specifically, the expanded circRNA-disease association adjacency matrix \mathbf{A}' is decomposed into two low-dimension matrices via optimizing the following objective function:

$$\min_{\mathbf{C}, \mathbf{D}} \|\mathbf{A}' - \mathbf{C}\mathbf{D}^T\|_F^2 \quad s.t. \quad \mathbf{C} \geq 0, \mathbf{D} \geq 0,$$

where \mathbf{C} is a latent feature matrix of circRNAs, \mathbf{D} is a latent feature matrix of diseases.

Furthermore, the algorithm constraints the latent feature spaces using two graph regularization terms and employs Frobenius norm regularization term to avoid overfitting and to enforce the smoothness of target space. The full objective function of the algorithm is formulated as follows:

$$\min_{\mathbf{C}, \mathbf{D}} \|\mathbf{A}' - \mathbf{C}\mathbf{D}^T\|_F^2 + \alpha \|\mathbf{C}\mathbf{D}^T\|_F^2 + \beta (\text{Tr}(\mathbf{C}^T \mathbf{G}^C \mathbf{C}) + \text{Tr}(\mathbf{D}^T \mathbf{G}^D \mathbf{D}))$$

$$s.t. \quad \mathbf{C} \geq 0, \mathbf{D} \geq 0,$$

where α and β are regularization coefficients. To solve for the low-dimension latent feature matrices \mathbf{C} and \mathbf{D} , Lagrange multipliers and Kuhn–Tucker conditions are utilized. The predicted circRNA-disease association scores are then calculated as:

$$\mathbf{A}^* = \mathbf{C}\mathbf{D}^T$$

■ 3.3.2 Deep learning-based methods

In recent years, deep learning methods, an important subgroup of machine learning, have emerged as powerful tools for discovering hidden patterns and associations in numerous areas, including the computational biology. This category is composed of models and techniques such as convolutional neural networks (CNN), recurrent neural networks (RNN), or DeepWalk. The methods based on deep learning category can be further divided into sub-categories, such as methods based on neural networks, methods based on graph learning and methods based on Markov graphs.

CRPGCN. In the IGNSCDA [38] algorithm from 2022, a graph convolutional network is designed to obtain the feature vectors of circRNAs and diseases, and based on them the multi-layer perceptron predicts circRNA-disease associations.

CRPGCN algorithm. The CRPGCN method [51] from 2021 uses a RWR algorithm to calculate the similarity between circRNAs and the similarity between diseases, PCA method is employed for dimensionality reduction and extracting features, and GCN algorithm is used for feature learning and calculating the final similarity scores between circRNAs and diseases.

THGNCD algorithm. The THGNCD algorithm [26] from 2023 constructs a triple heterogeneous graph network with nodes representing three different entities: circRNAs, diseases and miRNAs. The graph network integrates data from multiple sources, such as known circRNA-disease associations, circRNA-miRNA interactions, disease-miRNA interactions, disease semantic similarity based on disease classification from MeSH database, and circRNA sequence similarity based on the Levenshtein distance between circRNA sequences.

The processed data are used to generate a topological feature embedding for each circRNA and each disease, and an attribute feature embedding for each circRNA-disease pair, so that these embeddings capture the relationship of the entities with other circRNAs, diseases and miRNAs.

The embeddings are processed separately in two paths. The first path employs a graph neural network to extract information insights to extract topological features. It contains also a context attention layer for learning importance of node neighbours. The second path examines the attributes of circRNA-disease pairs through a multilayer convolutional neural network. The learnt representations from both paths are integrated to obtain output association scores of a circRNA-disease pair.

Wang's method. The method [70] firstly fuses multi-source information including disease semantic similarity, disease Gaussian interaction profile kernel similarity and circRNA Gaussian interaction profile kernel similarity. Secondly, a Convolutional Neural Network architecture is used for extraction of hidden deep features of the circRNA-disease relationships. The model of the architecture has a multi-layer neural network structure consisting of an input layer, a series of alternating convolution and subsampling layers,

finishing with full connection layer and the output layer. The feature map of the i th layer, denoted as L_i , is formulated as follows:

$$L_i = f(L_{i-1} \otimes W_i + b_i),$$

where $f(x)$ represents the activation function, L_{i-1} is the feature map of the $(i - 1)$ th layer, W_i represents the weight matrix of the convolution kernel of i th layer, b_i represents the offset vector and operator \otimes denotes a convolution operation. Convolution layers are followed by subsampling layers, which sample the feature graph. Lastly, an extreme learning machine (ELM), which is a learning algorithm based on single hidden layer feedforward neural network model, is used as a classifier to predict circRNA–disease associations. The ELM with h hidden layer nodes is formulated as follows:

$$o_j = \sum_{i=1}^h \beta_i f(W_i \cdot X_j + b_i) \quad j = 1, \dots, n,$$

where o_j represents the output for the sample with attribute X_j , β_i represents the output weight for the hidden node, $W_i \cdot X_j$ is the inner product of W_i and X_j , b_i is the offset of the i th hidden layer node, and n is the number of samples.

3.4 Related Problems

Within the field of bioinformatics, we can find a multitude of similar graph-based problems, such as *predicting protein-protein interactions*, *analyzing biological pathways*, *identifying connections between specific genes and diseases* or *drug-target interaction prediction*. Some of the even more closely related problem are *miRNA-disease association prediction* and *circRNA-miRNA interaction prediction*, as they both play an essential role in the very mechanism of circRNA-disease associations.

3.5 Summary

The presented methods approach the problem of circRNA-disease prediction from numerous and vastly different perspectives, sometimes integrating techniques from different categories together. Multiple of the presented methods come with several disadvantages, namely network-based and path-based methods are very dependent on the quality and completeness of the input interactions, for example KATZHCD algorithm performs poorly on sparse matrix, and these methods are biased towards well-studied entities. Furthermore, network-based methods may struggle to scale to large and complex networks, while path-based methods may not capture all the necessary biological information just based on paths. This work will in following chapters focus on graph neural network methods, as it can be still considered as a fairly novel approach, with a lot of unexplored ground.

The network-based and path-based methods come with several disadvantages, namely they are very dependent on the quality and completeness of the input interactions and are biased towards well-studied entities. In the study, KATZHCDA algorithm performs poorly on sparse matrices. Furthermore, network-based methods may struggle to scale to large and complex networks, while path-based methods may not capture all the necessary biological information just based on paths.

Chapter 4

Data Sources

4.1 Introduction

The exploration of circRNAs has brought deeper understanding of their diverse roles in cellular processes and disease pathogenesis. These accumulating findings motivated establishment of databases, both manually curated and computationally generated. In this chapter, several important databases amassing information about circRNA and disease associations will be introduced. Despite recognizing the extensive content of the following databases and tools, only the relevant functionalities and datasets will be explored in this work.

4.2 Databases

4.2.1 CircRNA-disease association Data Sources

CircR2Disease & CircR2Disease v2.0. Circ2Disease [82] is a manually curated database that holds experimentally supported circRNA-disease associations found in humans. Precisely, it contains 273 associations between 237 circRNAs and 54 human diseases with additional information, based on 120 studies. In 2022 circRNADisease v2.0 database [20] was released. As of current date, having reviewed 12,000 published literature, 6,998 associations were determined, involving 4,246 circRNAs and 330 standard diseases and 12 species.

CircR2Cancer. CircR2Cancer [39] is a manually curated database of associations between circRNAs and cancers. It contains 1439 experimentally verified circRNA-cancer relationships, including 1135 circRNAs and 82 cancers.

Circad. Database circad [56] a comprehensive manually curated database of circular RNAs associated with diseases. It lists 1388 circRNAs related to with 150 diseases by 1388 associations. The circRNAs span 4 species: human, mouse, rat and chicken.

CDASOR. CDASOR [50] is an algorithm that computes circRNA-disease association predictions based on sequence and ontology representations and convolutional and recurrent neural networks. The computed predictions are freely available.

4.2.2 CircRNA-miRNA Association Data Sources

One of the functions of circRNAs is post-transcriptional gene regulation by sponging miRNAs and binding RNA-binding proteins (RBPs), which then can no longer bind to an mRNA.

CircInteractome. CircInteractome (circRNA interactome) [16] web tool uses the TargetsCan prediction tool [25] to predict potential circRNA targets for miRNAs. The tool can also predict binding of circRNA with RBPs.

4.2.3 MiRNA-disease Association Data Sources

HMDD v4.0. The manually curated Human MicroRNA Disease Database v4.0 (HMDD v4.0) [14] holds 53530 miRNA-disease associations collected from published literature. New added associations now also include exosomal miRNAs and virus-encoded miRNAs.

4.2.4 Gene Related Data Sources

DisGeNET. DisGeNET contains associations of genes and variants with human diseases. DisGeNET v7.0 holds 1,134,942 gene-disease associations between 21,671 genes and 30,170 diseases. The database aggregates data from expert curated repositories, genome-wide association study catalogues, animal models and the scientific literature.

MiRDB. MicroRNAs are short noncoding RNAs that are involved in the regulation of gene expression of thousands of gene targets. Database miRDB [71] contains predicted gene targets computed by a bioinformatics tool MirTarget. The database, unlike others, primary focuses on mature miRNAs, which are the ones responsible for the miRNA-mediated gene expression regulation, and hosts predicted miRNA targets in five species: human, mouse, rat, dog and chicken.

4.2.5 Disease Related Data Sources

Disease Vocabularies. There are three widely used disease vocabularies: OMIM from the 1960s, MeSH from the 1960s, and DO [62] from the 2003, all serving different purposes. OMIM (Online Mendelian Inheritance in Man) serves as an online catalog of human genes and genetic disorders and traits. Medical Subject Headings (MeSH) is a comprehensive list of controlled vocabulary used for classification and indexing publications in the National Library of Medicine, with diseases being one of the categories of MeSH terms. Disease Ontology (DO), a standardized ontology for human disease, provides

descriptions of human disease terms, phenotype characteristics and related medical vocabularies. Across the different vocabularies, some terms have direct counterparts, some do not.

■ 4.2.6 CircRNA Related Data Sources

CircAtlas 3.0. CircAtlas 3.0 database [77] is a database of circRNAs and their expression and functional profiles in vertebrates. It contains over 3.1 million circRNAs across 10 species (human, macaque, mouse, rat, pig, chicken, dog, sheep, cat, rabbit) and various tissues, a rich collection of 2674 circRNA sequencing datasets, both Illumina and Nanopore. The database employs a standardized nomenclature scheme for circRNAs and also provides information of the host gene and circRNA exons.

■ 4.3 Challenges

There is an important question regarding the choice of the data sources - whether to use data only from manually curated databases with experimentally verified information or expand the certain information with the more imprecise computationally generated data. The balance between the two is important.

Chapter 5

Method

5.1 CircRNA-Disease Network Design

In order to let the model capture the intricacies of the circRNA-disease association, the graph has contain meaningful nodes and edges. Since the mechanism by which circRNAs relate to diseases often lies in binding miRNAs, which in turn bind to mRNAs, the addition of nodes representing miRNAs and mRNAs, i.e. the genes from which they are transcribed, to the graph is natural. As for the edges in the graph, denoting the interactions between the nodes, following associations were chosen:

- circRNA-disease
 - circRNAs bind miRNAs, which in turn bind to mRNAs
- circRNA-miRNA
 - circRNAs bind miRNAs
- miRNA-gene
 - miRNA regulates gene expression by binding to mRNAs transcribed from certain genes
- miRNA-disease
 - miRNA regulation of gene expression can lead to a disease
- disease-gene
 - abnormalities in gene expression can lead to a disease

The data for the interactions were taken from the data sources described by the table below. Regarding the circRNA-disease associations, it was opted for the CDASOR dataset, because, even though it is computationally generated, it contains the needed amount of associations.

The resulting network is provided in the figure below.

Association type	Database
circRNA – miRNA	circInteractome
circRNA – disease	CDASOR
disease – miRNA	HMDD v4.0
miRNA – gene	miRDB
disease – gene	DisGeNET
disease – disease	disease ontology

Table 5.1: A list of databases for various association types

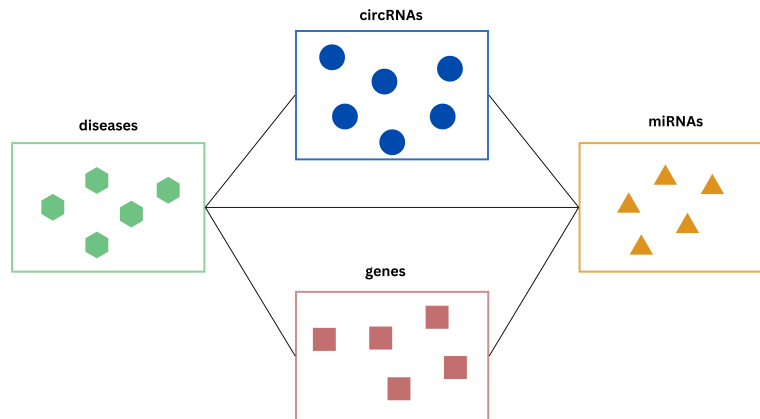


Figure 5.1: Illustration of the network design, edges are labelled with the data sources for the respective interactions

5.1.1 Data Preparation Challenges

1. Disease vocabularies

There are various popular disease nomenclature systems in use, leading to the diseases featured in different databases being documented under different names. While in some cases, the differences between aliases for one disease lie in capitalization of the first letters or in apostrophes, in other cases are parts of the names inconsistent, or even the whole disease names are lacking their equivalents in other disease vocabularies.

2. CircRNA naming

CircRNA naming is not just a problem of the era of circRNA discovery, but remains an unresolved issue even nowadays, as is explained in the article "A guide to naming eukaryotic circular RNAs" [5], which was published in 2023 under Nature. The article brings attention to the problem that circRNAs are often registered under different names across different databases. For example circRNA FAM120A is listed as HSA_CIRCpedia_64725 in database CIRCpedia, hsa-FAM120A_0006 in circAtlas, hsa_circFAM120A_007 in circBank, and hsa_circ_0001875

in circBase. They call on applying an intuitive naming scheme, which according to them should include the prefix 'circ', host gene symbol, and exon and intron information.

3. MiRNA naming

Names of miRNAs consist of a varying number of subparts connected by a dash, e.g. hsa-miR-1200. The first part made of three letters specifies the species. "hsa" stands for Homo sapiens. For the following part, to refer to a gene locus and precursor miRNA (i.e. pre-miRNA) of a miRNA, "mir" is used, while a mature miRNA product is referred to as "miR".

This has a consequence that effectively one miRNAs may be referred to by two distinct names: circInteractome keeps mature miRNAs (e.g. hsa-miR-1289), while HMDDv4 gene locuses (hsa-mir-1289).

MiRNAs with highly similar sequences, differing in a 1 or 2 nucleotides, as hsa-mir-151a and hsa-mir-151b, are differentiated by a suffix lower case letter.

Lastly, two diverse loci producing the identical mature miRNA are distinguished by an additional number separated by a dash, e.g. hsa-mir-125b-1 vs hsa-mir-125b-1.

5.1.2 Data Preprocessing

Names of miRNAs, diseases and circRNAs from different databases are respectively unified by following techniques:

- unifying diseases
 - curation of the disease name using the disease ontology IDs (DOIDs) and disease ontology terms with some manual assigning of DOID of diseases with the best match in the disease description and conversion between MeSH and DOID ontology identifiers when possible
- unifying circRNAs
 - searching for a match in aliases of circRNAs
- unifying miRNAs
 - not distinguishing between gene loci and mature product made from the gene by conversion of the name to lowercase

Afterwards, mappings are created for each entity type, that is circRNAs, diseases, miRNAs and genes, so that each unique entity name is mapped to a numerical id. A list of edges reflecting the new mappings is created for each edge type.

After the data are prepared, we create a HeteroData object and specify 4 types of nodes: circRNAs, miRNAs, diseases and genes, and initialize 5 types of edges between these node types, along with the reverse edges for each edge type. The resulting network is described by the tables below.

Edge Type	Count
circRNA – miRNA	168 841
circRNA – disease	3 221
disease – miRNA	53 530
miRNA – gene	465 741
disease – gene	5 963

Table 5.2: A list of edge types in the network and their counts

Node Type	Count
circRNA	5201
disease	2386
miRNA	3112
gene	17658

Table 5.3: A list of node types in the network and their counts

It is important look at the network that arises from the mixture of the five data sources. It can be seen that the amounts of different types of nodes and edges are highly imbalanced, what may later negatively affect the learning process. For example, the dominant miRNA connections with circRNAs might not be as important as the connections with diseases, and may therefore skew the circRNA node embeddding.

Below we can see a node degree distribution of circRNA nodes represented by a histogram. Most of the circRNA nodes have a low degree, with the highest degree being 317, the a circRNA node has on average 33.1 connections and the median node has 19 connections.

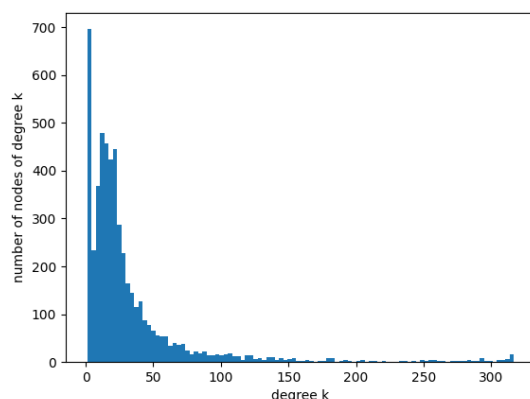


Figure 5.2: Node degree distribution of circRNA nodes

On the other hand, diseases have much more imbalanced connections. Around every fourth disease has just one connection, as can be seen on the graph of degree distribution of disease nodes with the logarithmic y axis below. On average, a disease node has 26.3 connections, but the median number of connections is only 3, meaning that diseases are very sparsely connected.

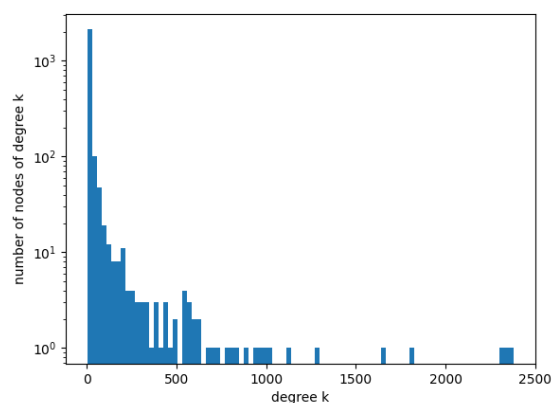


Figure 5.3: Node degree distribution of disease nodes

Regarding only circRNA-disease connections, a circRNA has on average 1.34 connections with a disease, or 1 connection in median, while a disease node has on average 22.2 connections with circRNAs or 5 connections in median.

■ 5.2 Graph Neural Network Model

For the task of link prediction on the graph, a graph neural network based on GraphSAGE framework was chosen, since it was outperforming other

architectures in study comparing link prediction models.

Firstly, an embedding layer is used to generate shallow embeddings for the initial feature representations for all nodes in the network, since the features are not specified explicitly, as all the information from data sources was used in the graph structure. Secondly, an encoder comprising of two layers of the GraphSAGE operator [28] separated by a ReLU unit and a dropout layer are used to generate node embeddings in a latent space. During the message passing, the GraphSAGE operator calculates the new node embedding value as a linear combination of the transformed current node representation and the mean of the neighbouring nodes representations:

$$x'_i = W_1 x_i + W_2 * \text{mean}_{j \in \mathbb{N}(i)} x_j \quad (5.1)$$

$$f_{ReLU}(x'_i) = \max(0, x'_i) \quad (5.2)$$

The decoder is defined as the inner product between the embeddings of diseases and circRNAs. Finally, to predict potential associations, the decoder output is fed into the sigmoid function:

$$\hat{y} = \sigma(z_{dis} z_{circ}^T) \quad (5.3)$$

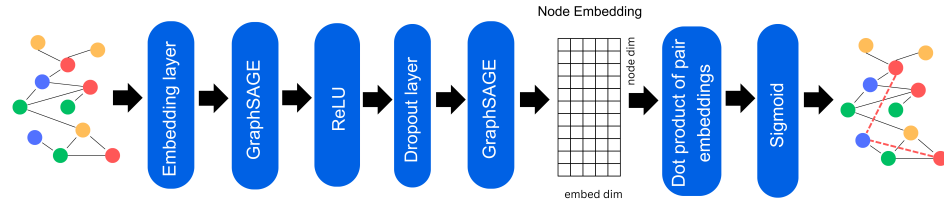


Figure 5.4: GNN model design

The design of the network can be seen in the provided figure below and the code for the model is provided in the appendix.

5.2.1 Loss function

The loss function assessing the quality of the trained embeddings is binary cross entropy.

$$BCE(y, \hat{y}) = -[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})],$$

where \hat{y} is the predicted probability and y the ground truth.

5.2.2 Implementation

The implementation of the proposed approach is built on top of the PyTorch Geometric, a library [24] designed for writing and training graph neural networks applicable in a wide range of problems dealing with structured

data. This popular library leverages the widespread adoption of PyTorch and offers various methods for deep learning on graphs working upon the message passing interface with message and update functions and multiple pooling operations. It is especially efficient on sparse graphs, as its dedicated GPU scatter and gather kernels can operate on all edges and nodes in parallel.

Chapter 6

Results and Discussion

6.1 Results

In order to evaluate the performance of the proposed graph neural network model, three different approaches, with each one illustrating a different perspective, were considered.

6.1.1 General Evaluation

The algorithm was trained and tested on ten different splits of the input data. The table below lists the evaluation metrics from the first five runs, and averages of the metrics after five and ten runs. The selected metrics are area under the curve (AUC), area under the precision-recall curve (AUPRC), accuracy (ACC), recall (REC) and precision (PRE).

When evaluating a binary classification task, we might have different intentions or goals, and therefore face trade-offs between sensitivity and specificity, or identifying correctly one class better at the expense of the other. In our case, we might want to either choose whether we put more importance on identifying more predictions even at the price of being incorrect, so that the possible associations can be then measured experimentally, or if want to focus on choosing only the most probable associations at the price of not identifying some, so that we can save money on doing unnecessary experiments with uncertain results. It could be said that the recall metric, which measures how often a model correctly identifies positive instances out of all the actual positive instances, reflects the first intention. On the other hand, the precision metric, which measures the proportion of true positives among positive predictions, is suitable for measuring for the second preference, as it measures how often are the positive predictions correct.

In cases of the imbalanced dataset, as happens also in this work, since the negative edges are samples with a ratio 2:1 to positive edges, it is recommended to use the precision-recall curve instead of the usual ROC curve (receiver operating characteristic curve). The precision-recall curve plots precision against recall, so the approach inherently focuses on the performance within the positive class, and the larger number of negative instances won't skew our perception of the model performance.

	Fold	AUC	AUPRC	ACC	REC	PRE
[H]	1	0.9464	0.8962	0.9555	0.9193	0.9457
	2	0.9146	0.8392	0.9296	0.8696	0.9150
	3	0.9340	0.8601	0.9410	0.9130	0.9102
	4	0.9503	0.8739	0.9493	0.9534	0.9003
	5	0.9092	0.8468	0.9306	0.8447	0.9412
	average (5)	0.9309	0.8632	0.9412	0.9000	0.9225
	average (10)	0.9304	0.8623	0.9408	0.8994	0.9220

Table 6.1: Results from five evaluation metrics: Area Under the Curve (AUC), Area Under the Precision-Recall Curve (AUPRC), Accuracy (ACC), Recall (REC), Precision (PRE)

As for the results themselves, it can be seen that the AUC is consistently above the 0.90 level, which is generally accepted as a very good performance, and is trailed by the AUCPRC metric, which is a little bit behind. The lower values for the AUPRC in comparison with AUC are expected, since it can be argued that it is easier to predict the negative class as the model is feeded more negative than positive edges. Accuracy can be explained analogously. As was discussed above, the AUCPR, most likely, better reflects the actual performance of the model than the AUC, with values just bellow 0.90.

Considering the two outlined possible intentions of the task, the second one (aiming for a high proportion of true positives among positive predictions), which is reflected by the precision metric, seems to be more successfully handled by the model than the first intention. The average precision hovers at 92% and the average recall at 90%.

6.1.2 Training Splits

The input data is split in a way that *circRNA-disease edges* in training data are not included in validation and test data. Afterwards, the *circRNA-disease edges* in training data are split once again: between edges designated for message passing and edges designated for supervision. Different ratios of splitting were tested to evaluate, whether it has any effect on the performance of the model. The results are displayed in the table below. Training split of 10% represents usage of 10% for supervision and 90% for message passing.

The results display only little differences between different ratios used for splitting the *circRNA-disease edges* in the training data for message passing and supervision. However, a weak pattern can be seen, as the performance across all metrics seems to increase with the increasing proportion of edges assigned to supervision. On one hand, this might possibly be a sign of overfitting, on the other hand, a higher amount of training data is always for the benefit of the model.

Training Split	AUC	AUPRC	ACC	REC	PRE
10%	0.9430	0.8935	0.9537	0.9109	0.9483
30%	0.9474	0.9071	0.9589	0.9130	0.9618
50%	0.9565	0.9140	0.9634	0.9358	0.9541
70%	0.9616	0.9206	0.9668	0.9461	0.9541
90%	0.9601	0.9198	0.9661	0.9420	0.9559

Table 6.2: Results from the training splits evaluation, evaluated with five evaluation metrics: area Under the Curve (AUC), area under the precision-recall curve (AUPRC), accuracy (ACC), recall (REC), precision (PRE)

6.1.3 Case Study

To further evaluate the performance of the model, a "case study" was conducted for a deeper analysis of the model predictions in a context. Given associations predicted as positive from a test dataset (scored by the dot product of the respective circRNA and disease embeddings) the top 10 predictions for false positives outcomes were examined further. False positives are outcomes where the model incorrectly predicts the positive class. The reason for this is to investigate the nature of such incorrect predictions – whether the circRNA-disease pairs are not related at all, which would mean that the model generates embeddings that do not reflect reality well, or whether the circRNA-disease pairs are at least somewhat similar. Also, the mechanisms behind the associations identified as genuine by studies outside of the input dataset are provided in the table below and described in a higher detail subsequently below it.

rank	disease	circRNA	mechanism
1	breast cancer	AFF2	miRNA sponging
2	leukemia (AML)	circAF4	no evidence
3	bladder carcinoma	circFARSA	miRNA sponging
4	colon cancer	circBC048201	no evidence
5	lung cancer	circ_CC SER1	no evidence
6	liver fibrosis	circBRIP	no evidence
7	cancer	hsa_circ_0000479	miRNA sponging
8	bladder carcinoma	circ_CDR1	miRNA sponging
9	lung cancer	circIPO11	no evidence
10	brain ischemia	circ_ERC1	no evidence

Table 6.3: A ranked list of associations predicted by the model that were marked as false positives based on the input dataset

The predicted associations listed in the table are further examined below, following the same order of appearance as in the table:

1. breast cancer – AFF2: circ-AFF2 can sponge miRNA-638, which can affect chemoresistance in breast cancer patients [89]

2. acute myeloid leukemia – circAF4: while it was shown that circRNA circAF4 regulates MLL-AF4 fusion protein expression and thus can inhibit mixed lineage leukemia progression [29], there is not a documented link to acute myeloid leukemia
3. bladder carcinoma – circFARSA: circFARSA expression is upregulated and has oncogenic functions in bladder cancer by sponging miR-330 [23]
4. colon cancer – circBC048201: paper [33] describes how circBC048201 acts as a sponge for miR-1184, while a separate study [6] documents how the miR-1184 regulates the proliferation and apoptosis of colon cancer cells
5. lung cancer – circ_CCSE1: no evidence
6. liver fibrosis – circBRIP: study [22] discovers that circBRIP performed well at distinguishing NSCLC lung cancer from benign pulmonary diseases, but a link to liver fibrosis was not found
7. cancer – hsa_circ_0000479: hsa_circ_0000479 promotes tumor progression by sponging miR-370-3p which regulates MSH2 expression [75] and also regulates ovarian cancer progression via sponging miR-942, which in turn regulates EPSTI1 expression [76]
8. bladder carcinoma – circ_CDR1 Cdr1as sensitizes bladder cancer to cisplatin by upregulation of APAF1 expression by sponging miR-1270 [84]
9. lung cancer – circIPO11: the circIPO11 has been only linked to other types of cancer, namely colorectal cancer [54] and liver cancer [55]
10. brain ischemia – circ_ERC1: no evidence was found for this pair, however, study [88] found that a circRNA from a Erc1 gene is dysregulated in rat brains during intracerebral hemorrhage, which is another type of stroke, but the two conditions are interconnected as hemorrhagic events can lead to ischemic damage

From the top 10 predicted associations marked as false positives, 4 were directly found to be proved by studies outside of the input data, and another 5 were found to be plausible candidates, since either associations with closely related diseases has been proved, or separate studies were found for the circRNA-miRNA interaction and miRNA-disease association. It can be concluded that the model manages to predict novel (not seen in the input data) associations between circRNAs and diseases.

6.2 Discussion

This section will assess the overall performance of the proposed model and general suitability of the graph neural network approaches for the task of

predicting circRNA-disease associations, and outline possibilities for future research in this topic.

6.2.1 Evaluation of the Proposed Model

Regarding the implemented model, based on the provided evaluation, it can be concluded that it performs well in the assigned task, reaching AUCPRC of 86% on average. Its main contributions lie in that firstly, it integrates numerous interactions, namely circRNA – miRNA, circRNA – disease, disease – miRNA, miRNA – gene and disease – gene. These interactions and associations are used for the graph construction itself, so the information encoded in them is not distorted by any pre-processing technique. Secondly, it employs the graphSAGE operator, which has not been used in any of the found related work on circRNA-disease association prediction, but outperforms other operators on general link prediction tasks. Thirdly, it is built upon the Pytorch Geometric Framework, meaning it is easily customizable and flexible, as well as optimized for efficient graph operations.

6.2.2 Limitations of the Model

The model presented in this work has multiple limitations.

Firstly, this work views the problem at hand under the supervised training setting. However, this setting does not really fit the problem, since the training data samples are in fact positive & unlabeled rather than positive & negative. For this reason, the approach in this work inherently leads to mislabelling. While there are some available graph neural network approaches for positive unlabeled learning for graph data, such as a method based on a GAT developed by Wu et al. from 2021 [74], these methods deal with unlabeled nodes, with the research regarding unlabeled links remaining very limited.

Secondly, the model does not use any explicit feature vectors for the initialization of the hidden states of nodes, since all the information from the datasets was utilized for building the graph itself. Additional possibilities, were considered, but found lacking in one way or another. For example, calculation of similarities between pairs of circRNAs would be suitable for interconnecting the most similar circRNAs together. However, the existing methods either employed Levenshtein distance between two circRNA sequences for measuring their similarities, which is inherently flawed since the most of the circRNA sequences has no effect on their interactions, and it is the miRNA binding sites that make all the difference; or other methods calculated similarities between circRNAs based on their interaction profiles with diseases, relying on the premise that similar circRNAs interact with similar diseases. The notion can be presumed to be based on the fact that similar circRNAs will contain the same binding sites and will therefore interact with the same diseases, which us consequently once again leads to the fact, that miRNA binding sites and interactions with miRNAs are crucial knowledge.

Thirdly, GraphSAGE operator itself has a few limitations. It predominantly captures local structural information, therefore it will miss global structural patterns. Furthermore, the number of sampled neighbours is a static parameter, which might not be ideal when we have some nodes with many neighbours and some with only few neighbours, as in our case.


Fourthly, inherently, the performance of the model is heavily influenced by the known circRNA-disease associations. Incorrect associations in the training data may distort the real patterns that we are looking for.

■ 6.2.3 Future Work

In future work, more sources of biological information could be incorporated into the edges of the graph as well as the feature vectors of the nodes for the initialization of the hidden states, for example, circRNAs similarities and disease similarities. For a disease similarity measure, Wang’s method [10] could be used, which is based on the hierarchical structure of a disease ontology. For the circRNA similarities could be calculated from the circRNA expression profiles.

Furthermore, methods for positive unlabelled learning based on graph neural networks, such as [74], should be investigated regarding its possible application for not just this problem, but for the link prediction tasks within the field of bioinformatics.

Lastly, taking into account possible incorporation of new data, a graph attention module could be incorporated for the purpose of assigning different weights to different neighbours.



Chapter 7

Conclusion

CircRNAs are small molecules that have been found to be dysregulated during multitude of diseases and thus show potential as disease biomarkers or therapeutic targets. Exploring associations between circRNAs and diseases experimentally is expensive and time-consuming, and therefore the computational approaches are employed in the task. The purpose of this thesis was to investigate the application of graph neural network on the problem of predicting these circRNA-disease associations.

The work on the thesis was broken down into several steps, which can be also seen to be reflected in the composition of the thesis. Firstly, the biological meaning and mechanisms behind the circRNA-disease associations were inquired into, since such knowledge is imperative for understanding and design of the methods employed in solving the problem. Secondly, having studied the graph neural networks, the main concepts and essential models were introduced in the second chapter of the thesis. Thirdly, existing methods, which fall into three large categories: network propagation-based methods, path-based methods, and machine learning methods, have been explored and representative methods from the respective categories were presented. After that, it was searched for data sources containing relevant data about interactions and characterization of circRNA and diseases.

Having collected all the necessary information, the problem was formulated as a link prediction task on a graph and a graph neural network was designed for association prediction. The model based on the graphSAGE framework was implemented upon Pytorch Geometric library.

The evaluation of the model performance shows that the model reaches the AUC of 0.92 and the AUCPR of 0.86 on average. It is important to mention that the model also managed to predict associations that were outside of information included in the input data.

The performance of the model could be further improved with the addition of other biological information, such as knowledge of similarities between different circRNAs and similarities between different diseases themselves. Another proposed improvement for the future work is to employ positive unlabelled learning based on graph neural networks.



Bibliography

- [1] Djihaad Arrar, Nadjat Kamel, and Abdelaziz Lakhffif. A comprehensive survey of link prediction methods. *The Journal of Supercomputing*, 80(3):3902–3942, Feb 2024.
- [2] Junwen Chen, Jun Yang, Xiang Fei, Xia Wang, and Kefeng Wang. CircRNA ciRS-7: A novel oncogene in multiple cancers. *Int. J. Biol. Sci.*, 17(1):379–389, January 2021.
- [3] L Chen, S Zhang, J Wu, J Cui, L Zhong, L Zeng, and S Ge. circRNA_100290 plays a role in oral cancer by functioning as a sponge of the mir-29 family. *Oncogene*, 36(32):4551–4561, August 2017.
- [4] Liang Chen, Chuan Huang, Xiaolin Wang, and Ge Shan. Circular RNAs in eukaryotic cells. *Curr. Genomics*, 16(5):312–318, July 2015.
- [5] Ling-Ling Chen, Albrecht Bindereif, Irene Bozzoni, Howard Chang, A. Matera, Myriam Gorospe, Thomas Hansen, Jørgen Kjems, Xu-Kai Ma, Jun Pek, Nikolaus Rajewsky, Julia Salzman, Jeremy Wilusz, Li Yang, and Fangqing Zhao. A guide to naming eukaryotic circular rnas. *Nature Cell Biology*, 25, 01 2023.
- [6] Shuo Chen, Yan Wang, Mingyue Xu, Lin Zhang, Yinan Su, Boxue Wang, and Xipeng Zhang. mir-1184 regulates the proliferation and apoptosis of colon cancer cells via targeting CSNK2A1. *Mol. Cell. Probes*, 53(101625):101625, October 2020.
- [7] Xiaoping Chen, Ping Han, Tao Zhou, Xuejiang Guo, Xiaofeng Song, and Yan Li. circrnadb: A comprehensive database for human circular rnas with protein-coding annotations. *Scientific Reports*, 6(1):34985, Oct 2016.
- [8] Xue-Jiao Chen, Zi-Cheng Zhang, Xiao-Yun Wang, Heng-Qiang Zhao, Meng-Lan Li, Yue Ma, Yang-Yang Ji, Chang-Jun Zhang, Kun-Chao Wu, Lue Xiang, Lan-Fang Sun, Meng Zhou, and Zi-Bing Jin. The circular RNome of developmental retina in mice. *Mol. Ther. Nucleic Acids*, 19:339–349, March 2020.

- [9] Chenglong Cheng, Yuting Wang, Qiuyun Xue, Yurong Huang, Xiao Wang, Faxue Liao, and Chenggui Miao. CircRnas in atherosclerosis, with special emphasis on the spongy effect of circrnas on mirnas. *Cell Cycle*, 22(5):527–541, March 2023.
- [10] Liang Cheng, Hengqiang Zhao, Pingping Wang, Wenyang Zhou, Meng Luo, Tianxin Li, Junwei Han, Shulin Liu, and Qinghua Jiang. Computational methods for identifying similar diseases. *Mol. Ther. Nucleic Acids*, 18:590–604, December 2019.
- [11] Yuning Cheng, Si-Mei Xu, Konii Takenaka, Grace Lindner, Ashton Curry-Hyde, and Michael Janitz. A unique circular rna expression pattern in the peripheral blood of myalgic encephalomyelitis/chronic fatigue syndrome patients. *Gene*, 877:147568, 2023.
- [12] Matthew Cobb. 60 years ago, francis crick changed the logic of biology. *PLOS Biology*, 15(9):1–8, 09 2017.
- [13] Claude Cocquerelle, B Mascrez, D Héтуin, and B Bailleul. Mis-splicing yields circular rna molecules. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 7:155–60, 02 1993.
- [14] Chunmei Cui, Bitao Zhong, Rui Fan, and Qinghua Cui. HMDD v4.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Research*, 52(D1):D1327–D1332, 08 2023.
- [15] Miri Danan, Schraga Schwartz, Sarit Edelheit, and Rotem Sorek. Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Research*, 40(7):3131–3142, 12 2011.
- [16] Ioannis Grammatikakis Supriyo De Kotb Abdelmohsen Dawood B. Dudekula, Amaresh C. Panda and Myriam Gorospe. Circinteractome: A web tool for exploring circular rnas and their interacting proteins and micrnas. *RNA Biology*, 13(1):34–42, 2016. PMID: 26669964.
- [17] Yulian Ding, Bolin Chen, Xiujuan Lei, Bo Liao, and Fang-Xiang Wu. Predicting novel circrna-disease associations based on random walk and logistic regression model. *Computational Biology and Chemistry*, 87:107287, 2020.
- [18] Mihnea Dragomir and George A. Calin. Circular rnas in cancer – lessons learned from micrnas. *Frontiers in Oncology*, 8, 2018.
- [19] Yehoshua Enuka, Mattia Lauriola, Morris E. Feldman, Aldema Sas-Chen, Igor Ulitsky, and Yosef Yarden. Circular RNAs are long-lived and display only minimal early alterations in response to a growth factor. *Nucleic Acids Research*, 44(3):1370–1383, 12 2015.

- [20] Chunyan Fan, Xiujuan Lei, Jiaojiao Tie, Yuchen Zhang, Fang-Xiang Wu, and Yi Pan. Circr2disease v2.0: An updated web server for experimentally validated circrna–disease associations and its application. *Genomics, Proteomics Bioinformatics*, 20(3):435–445, 2022. Bioinformatics Commons—2022.
- [21] Chunyan Fan, Xiujuan Lei, and Fang-Xiang Wu. Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks. *Int. J. Biol. Sci.*, 14(14):1950–1959, 2018.
- [22] Xinfeng Fan, Qi Zhang, Shiyi Qin, and Shaoqing Ju. CircBRIP1: a plasma diagnostic marker for non-small-cell lung cancer. *J. Cancer Res. Clin. Oncol.*, 150(2):83, February 2024.
- [23] Chen Fang, Xin Huang, Jun Dai, Wei He, Le Xu, and Fukang Sun. The circular RNA circFARSA sponges microRNA-330-5p in tumor cells with bladder cancer phenotype. *BMC Cancer*, 22(1):373, April 2022.
- [24] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. 2019.
- [25] Andrew Grimson, Kyle Kai-How Farh, Wendy K. Johnston, Philip Garrett-Engele, Lee P. Lim, and David P. Bartel. MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. *Molecular Cell*, 27(1):91–105, 2007.
- [26] Yuwei Guo and Ming Yi. THGNCA: circRNA–disease association prediction based on triple heterogeneous graph network. *Briefings in Functional Genomics*, page elad042, 09 2023.
- [27] William L. Hamilton. *The Graph Neural Network Model*, pages 51–70. Springer International Publishing, Cham, 2020.
- [28] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs, 2018.
- [29] Wei Huang, Ke Fang, Tian-Qi Chen, Zhan-Cheng Zeng, Yu-Meng Sun, Cai Han, Lin-Yu Sun, Zhen-Hua Chen, Qian-Qian Yang, Qi Pan, Xue-Qun Luo, Wen-Tao Wang, and Yue-Qin Chen. circRNA circAF4 functions as an oncogene to regulate MLL-AF4 fusion protein expression and inhibit MLL leukemia progression. *J. Hematol. Oncol.*, 12(1):103, October 2019.
- [30] Nicky Hudson. The missed disease? endometriosis as an example of ‘undone science’. *Reprod. Biomed. Soc. Online*, 14:20–27, March 2022.
- [31] William Jeck, Jessica Sorrentino, Kai Wang, Michael Slevin, Christin Burd, Jinze Liu, William Marzluff, and Norman Sharpless. Circular rnas are abundant, conserved, and associated with alu repeats. *RNA*, 19:426–426, 03 2013.

- [45] Shanbao Li, Junyong Weng, Fangbin Song, Lei Li, Chao Xiao, Weiqiang Yang, and Junming Xu. Circular RNA circZNF566 promotes hepatocellular carcinoma progression by sponging mir-4738-3p and regulating TDO2 expression. *Cell Death Dis.*, 11(6):452, June 2020.
- [46] Yan Li, Qiupeng Zheng, Chunyang Bao, Shuyi Li, Weijie Guo, Jiang Zhao, Di Chen, Jianren Gu, Xianghuo He, and Shenglin Huang. Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Research*, 25(8):981–984, August 2015.
- [47] Zhaoyong Li, Chuan Huang, Chun Bao, Liang Chen, Mei Lin, Xiaolin Wang, Guolin Zhong, Bin Yu, Wanchen Hu, Limin Dai, Pengfei Zhu, Zhaoxia Chang, Qingfa Wu, Yi Zhao, Ya Jia, Ping Xu, Huijie Liu, and Ge Shan. Exon-intron circular rnas regulate transcription in the nucleus. *Nature Structural & Molecular Biology*, 22(3):256–264, Mar 2015.
- [48] Zhouxiao Li, Ye Cheng, Fan Wu, Liangliang Wu, Hongyong Cao, Qian Wang, and Weiwei Tang. The emerging landscape of circular RNAs in immunity: breakthroughs and challenges. *Biomarker Research*, 8(1):25, July 2020.
- [49] Xianzhi Lin, Hsien-Chun Lo, David T. W. Wong, and Xinshu Xiao. Noncoding rnas in human saliva as potential disease biomarkers. *Frontiers in Genetics*, 6, 2015.
- [50] Chengqian Lu, Min Zeng, Fang-Xiang Wu, Min Li, and Jianxin Wang. Improving circRNA-disease association prediction by sequence and ontology representations with convolutional and recurrent neural networks. *Bioinformatics*, 36(24):5656–5664, April 2021.
- [51] Zhihao Ma, Zhufang Kuang, and Lei Deng. CRPGCN: predicting circRNA-disease associations using graph convolutional network based on heterogeneous network. *BMC Bioinformatics*, 22(1):551, November 2021.
- [52] Sebastian Memczak, Marvin Jens, Antigoni Elefsinioti, Francesca Torti, Janna Krueger, Agnieszka Rybak, Luisa Maier, Sebastian D Mackowiak, Lea H Gregersen, Mathias Munschauer, Alexander Loewer, Ulrike Ziebold, Markus Landthaler, Christine Kocks, Ferdinand le Noble, and Nikolaus Rajewsky. Circular rnas are a large class of animal rnas with regulatory potency. *Nature*, 495(7441):333–338, March 2013.
- [53] Sebastian Memczak, Panagiotis Papavasileiou, Oliver Peters, and Nikolaus Rajewsky. Identification and characterization of circular RNAs as a new class of putative biomarkers in human blood. *PLoS One*, 10(10):e0141214, October 2015.
- [54] Maria Radanova, Galya Mihaylova, Oskan Tasinov, Desislava P Ivanova, George St Stoyanov, Neshe Nazifova-Tasinova, Rostislav Manev, Ayshe Salim, Miglena Nikolova, Diana G Ivanova, Nikolay Conev, Zhasmina

- [64] Yanggu Shi and Jindong Shang. Circular rna expression profiling by microarraydash;a technical and practical perspective. *Biomolecules*, 13(4), 2023.
- [65] Yanggu Shi and Jindong Shang. Circular rna expression profiling by microarraydash;a technical and practical perspective. *Biomolecules*, 13(4), 2023.
- [66] Natalia G. Starostina, Sarah Marshburn, L. Steven Johnson, Sean R. Eddy, Rebecca M. Terns, and Michael P. Terns. Circular box c/d rnas in *pyrococcus furiosus*. *Proceedings of the National Academy of Sciences*, 101(39):14097–14101, 2004.
- [67] Laurent F. Thomas and Pål Sætrom. Circular RNAs are depleted of polymorphisms at microRNA binding sites. *Bioinformatics*, 30(16):2243–2246, 04 2014.
- [68] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.
- [69] Giovanni Visonà, Emmanuelle Bouzigon, Florence Demenais, and Gabriele Schweikert. Network propagation for GWAS analysis: a practical guide to leveraging molecular networks for disease gene discovery. *Briefings in Bioinformatics*, 25(2):bbae014, 02 2024.
- [70] Lei Wang, Zhu-Hong You, Yu-An Huang, De-Shuang Huang, and Keith C C Chan. An efficient approach based on multi-sources information to predict circRNA-disease associations using deep convolutional neural network. *Bioinformatics*, 36(13):4038–4046, July 2020.
- [71] Xiaowei Wang. miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA*, 14(6):1012–1017, June 2008.
- [72] Hang Wei and Bin Liu. iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. *Brief. Bioinform.*, 21(4):1356–1367, July 2020.
- [73] Wenhao Weng, Qing Wei, Shusuke Toden, Kazuhiro Yoshida, Takeshi Nagasaka, Toshiyoshi Fujiwara, Sanjun Cai, Huanlong Qin, Yanlei Ma, and Ajay Goel. Circular RNA ciRS-7—A promising prognostic biomarker and a potential therapeutic target in colorectal cancer. *Clin. Cancer Res.*, 23(14):3918–3928, July 2017.
- [74] Man Wu, Shirui Pan, Lan Du, and Xingquan Zhu. Learning graph neural networks with positive and unlabeled nodes. *ACM Trans. Knowl. Discov. Data*, 15(6), jun 2021.
- [75] Peng Wu, Jing Qin, Lingyan Liu, Wupeng Tan, Linchen Lei, and Jiayu Zhu. circEPSTII1 promotes tumor progression and cisplatin resistance via

- [86] Meng Zhang and Zhigang Bian. The emerging role of circular rnas in alzheimer’s disease and parkinson’s disease. *Frontiers in Aging Neuroscience*, 13, 2021.
- [87] Yang Zhang, Xiao-Ou Zhang, Tian Chen, Jian-Feng Xiang, Qing-Fei Yin, Yu-Hang Xing, Shanshan Zhu, Li Yang, and Ling-Ling Chen. Circular intronic long noncoding RNAs. *Mol. Cell*, 51(6):792–806, September 2013.
- [88] Yulan Zhong, Xiaoqiang Li, Chuqiao Li, Yudi Li, Yuqi He, Fangming Li, and Li Ling. Intracerebral hemorrhage alters circular RNA expression profiles in the rat brain. *Am. J. Transl. Res.*, 12(8):4160–4174, August 2020.
- [89] Fan Zhou, Dongjiao Wang, Wei Wei, Haimin Chen, Haotian Shi, Nian Zhou, Lixia Wu, and Rong Peng. Comprehensive profiling of circular RNA expressions reveals potential diagnostic and prognostic biomarkers in multiple myeloma. *BMC Cancer*, 20(1):40, January 2020.
- [90] Qian Zhou, Lin-Ling Ju, Xiang Ji, Ya-Li Cao, Jian-Guo Shao, and Lin Chen. Plasma circRNAs as biomarkers in cancer. *Cancer Manag. Res.*, 13:7325–7337, September 2021.
- [91] Yutian Zou, Shaoquan Zheng, Xinpei Deng, Anli Yang, Xinhua Xie, Hailin Tang, and Xiaoming Xie. The role of circular RNA CDR1as/ciRS-7 in regulating tumor microenvironment: A pan-cancer analysis. *Biomolecules*, 9(9):429, August 2019.

Appendix A

GNN Model

```
class GNN(torch.nn.Module):
    def __init__(self, hidden_channels_f, hidden_channels_s,
                 output_channels, dropout=0.2):
        super().__init__()

        self.conv1 = SAGEConv(hidden_channels_f, hidden_channels_s)
        self.conv2 = SAGEConv(hidden_channels_s, output_channels)
        self.dropout = dropout

    def forward(self, x: Tensor, edge_index: Tensor) -> Tensor:
        x = F.relu(self.conv1(x, edge_index))
        x = F.dropout(x, p=self.dropout)
        x = self.conv2(x, edge_index)
        return x

# the classifier applies the dot-product between source and
# destination node embeddings to derive edge-level predictions:
class Classifier(torch.nn.Module):
    def forward(self, x_circRNAs: Tensor, x_diseases: Tensor,
                edge_label_index: Tensor) -> Tensor:
        # Convert node embeddings to edge-level representations:
        edge_feat_circRNA = x_circRNAs[edge_label_index[0]]
        edge_feat_disease = x_diseases[edge_label_index[1]]

        # Apply dot-product to get a prediction per supervision edge:
        return (edge_feat_circRNA * edge_feat_disease).sum(dim=-1)

class Model(torch.nn.Module):
    def __init__(self, hidden_channels_f, hidden_channels_s,
                 output_channels):
        super().__init__()
        self.circRNA_emb =
            torch.nn.Embedding(data["circRNA"].num_nodes,
                               hidden_channels_f)
        self.disease_emb =
            torch.nn.Embedding(data["disease"].num_nodes,
```

```
        hidden_channels_f)
self.miRNA_emb = torch.nn.Embedding(data["miRNA"].num_nodes,
        hidden_channels_f)
self.gene_emb = torch.nn.Embedding(data["gene"].num_nodes,
        hidden_channels_f)

self.gnn = GNN(hidden_channels_f, hidden_channels_s,
        output_channels)

self.gnn = to_hetero(self.gnn, metadata=data.metadata())

self.classifier = Classifier()

def forward(self, data: HeteroData) -> Tensor:
    x_dict = {
        "circRNA": self.circRNA_emb(data["circRNA"].node_id),
        "disease": self.disease_emb(data["disease"].node_id),
        "miRNA": self.miRNA_emb(data["miRNA"].node_id),
        "gene": self.gene_emb(data["gene"].node_id)
    }

    # 'x_dict' holds feature matrices of all node types
    # 'edge_index_dict' holds all edge indices of all edge types
    x_dict = self.gnn(x_dict, data.edge_index_dict)
    pred = self.classifier(
        x_dict["circRNA"],
        x_dict["disease"],
        data["circRNA", "acd", "disease"].edge_label_index,
    )

    return pred
```
