Joint learning from multiple information sources for biological problems

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades

DOKTOR DER NATURWISSENSCHAFTEN
Dr. rer. nat.

genehmigte Dissertation
von
M. Sc. Thi Ngan Dong
geboren am 24. Juli 1988, in Hanoi, Vietnam

2022

Referent: Prof. Dr. -Ing Wolfgang Nejdl

Korreferentin: Assist. Prof. Megha Khosla

Tag der Promotion: 13.07.2023

Declaration of Authorship

I, Thi Ngan DONG, declare that this thesis titled, "Joint learning from multiple information sources for biological problems" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 13.07.2023

"What is not started will never get finished."

Johann Wolfgang von Goethe

ABSTRACT

Thanks to technological advancements, more and more biological data have been generated in recent years. Data availability offers unprecedented opportunities to look at the same problem from multiple aspects. It also unveils a more global view of the problem that takes into account the intricated inter-play between the involved molecules/entities. Nevertheless, biological datasets are biased, limited in quantity, and contain many false-positive samples. Such challenges often drastically downgrade the performance of a predictive model on unseen data and, thus, limit its applicability in real biological studies.

Human learning is a multi-stage process in which we usually start with simple things. Through the accumulated knowledge over time, our cognition ability extends to more complex concepts. Children learn to speak simple words before being able to formulate sentences. Similarly, being able to speak correct sentences supports our learning to speak correct and meaningful paragraphs, etc. Generally, knowledge acquired from related learning tasks would help boost our learning capability in the current task. Motivated by such a phenomenon, in this thesis, we study supervised machine learning models for bioinformatics problems that can improve their performance through exploiting multiple related knowledge sources. More specifically, we concern with ways to enrich the supervised models' knowledge base with publicly available related data to enhance the computational models' prediction performance.

Our work shares commonality with existing works in multimodal learning, multi-task learning, and transfer learning. Nevertheless, there are certain differences in some cases. Besides the proposed architectures, we present large-scale experiment setups with consensus evaluation metrics along with the creation and release of large datasets to showcase our approaches' superiority. Moreover, we add case studies with detailed analyses in which we place no simplified assumptions to demonstrate the systems' utilities in realistic application scenarios. Finally, we develop and make available an easy-to-use website for non-expert users to query the model's generated prediction results to facilitate field experts' assessments and adaptation. We believe that our work serves as one of the first steps in bridging the gap between "Computer Science" and "Biology" that will open a new era of fruitful collaboration between computer scientists and biological field experts.
Keywords: joint learning, learning from multiple sources, data integration, biological problems

ZUSAMMENFASSUNG

Dank des technologischen Fortschritts sind in den letzten Jahren immer mehr biologische Daten erzeugt worden. Die Datenverfügbarkeit bietet nie dagewesene Möglichkeiten, ein und dasselbe Problem aus verschiedenen Blickwinkeln zu betrachten. Sie ermöglicht auch eine umfassendere Sicht auf das Problem, die das komplizierte Zusammenspiel zwischen den beteiligten Molekülen/-Einheiten berücksichtigt. Dennoch sind biologische Datensätze voreingenommen, in ihrer Menge begrenzt und enthalten viele falsch-positive Proben. Derartige Herausforderungen verschlechtern die Leistung eines Vorhersagemodells bei ungesehenen Daten oft drastisch und schränken somit seine Anwendbarkeit in realen biologischen Studien ein.

Das menschliche Lernen ist ein mehrstufiger Prozess, bei dem wir normalerweise mit einfachen Dingen beginnen. Durch das im Laufe der Zeit angesammelte Wissen erweitert sich unsere Wahrnehmungsfähigkeit auf komplexere Konzepte. Kinder lernen, einfache Wörter zu sprechen, bevor sie in der Lage sind, Sätze zu formulieren. In ähnlicher Weise unterstützt die Fähigkeit, korrekte Sätze zu sprechen, unser Lernen, korrekte und sinnvolle Absätze zu formulieren usw. Im Allgemeinen trägt das bei verwandten Lernaufgaben erworbene Wissen dazu bei, unsere Lernfähigkeit bei der aktuellen Aufgabe zu steigern. Motiviert durch dieses Phänomen, untersuchen wir in dieser Arbeit überwachte maschinelle Lernmodelle für bioinformatische Probleme, die ihre Leistung durch die Nutzung mehrerer verwandter Wissensquellen verbessern können. Genauer gesagt beschäftigen wir uns damit, wie die Wissensbasis der überwachten Modelle mit öffentlich verfügbaren verwandten Daten angereichert werden kann, um die Vorhersageleistung der Computermodelle zu verbessern.

Unsere Arbeit weist Gemeinsamkeiten mit bestehenden Arbeiten im Bereich des multimodalen Lernens, des Multi-Task-Lernens und des Transfer-Lernens auf. Dennoch gibt es in einigen Fällen gewisse Unterschiede. Neben den vorgeschlagenen Architekturen präsentieren wir groß angelegte Experimente mit konsensuellen Bewertungsmaßstäben sowie die Erstellung und Freigabe großer Datensätze, um die Überlegenheit unserer Ansätze zu demonstrieren. Darüber hinaus fügen wir Fallstudien mit detaillierten Analysen hinzu, in denen wir keine vereinfachten Annahmen treffen, um die Nützlichkeit der Systeme in realistischen Anwendungsszenarien zu demonstrieren. Schließlich entwickeln wir eine einfach zu bedienende Website, auf der Nicht-Experten die vom Modell generierten Vorhersageergebnisse abfragen können, um die Bewertung und Anpassung durch Experten vor Ort zu erleichtern. Wir glauben, dass unsere Arbeit einer der ersten Schritte zur Überbrückung der Kluft zwischen “Informatik” und “Biologie” ist, die eine neue Ära der fruchtbaren Zusammenarbeit zwischen Informatikern und Biologen einleiten wird.

Schlagwörter: gemeinsames Lernen, Lernen aus mehreren Quellen, Datenintegration, biologische Probleme

Acknowledgements

I sincerely thank prof. Wolfgang Nejdl for giving me the PhD admission and supporting me throughout my PhD life. I wholeheartedly thank my supervisor, Megha Khosla, for being my devoted mentor, consultant, supporter, and critic. I want to thank my collaborators in the PRESENt, Big data for Cochlear Implants, and LeibnizKILabor projects for enlightening me with their expertise and guiding me in the right directions.

Finally, my thesis would not be possible without my family, friends, and colleagues. Thank you all for your endless support and for being with me through both hard and cheerful times.

Contents

List of Figures

List of Tables

List of Abbreviations

ACC	Accuracy
ACE2	Angiotensin-Converting Enzyme 2
ANPEP	Aminopeptidase N
AP	Area under the Precision-Recall curve
AUC	Area under Receiver Operating Characteristic curve
DAG	Disease directed acyclic graph
DDP4	Dipeptidyl peptidase 4
DG	Disease Gaussian Interaction Profile kernel similarity
DNA	Deoxyribonucleic acid
DP	Disease target similarity
DS	Disease semantic similarity
DT	Disease target gene feature
$\|E^+\|$	Number of positive and negative interactions
$\|E^-\|$	Number of positive and negative interactions
GIP	Gaussian Interaction Profile kernel
GO	Gene Ontology
HH	Human-human protein-protein interaction training set
LOSO	Leave-One-Species-Out
LSTM	Long Short Term Memory
MA	MiRNA sequence alighment similarity
MCC	Matthews correlation coefficient
MF	MiRNA functional similarity
MG	MiRNA GIP similarity
miRNA	Micro RNA
mLSTM	Multiplicative Long Short Term Memory
ML	Machine Learning
MLP	Multilayer Perceptrons
MO	MiRNA family feature
MP	MiRNA target similarity
MPM	Message passing framework for miRNA-Disease association prediction
mRNA	Messenger RNA
MS	MiRNA sequence geometric similarity
MT	MiRNA target gene feature
MTT	Multitask Transfer
MuCoMiD	Multitask learning framework for miRNA-disease association prediction
nt	Neucleotide
n_m	Number of miRNAs
n_d	Number of diseases
n_p	Number of PCGs

n_{md}	Number of miRNA-disease associations
n_f	Number of miRNA families
PCG	Protein-coding gene
PPI	Protein-protein interaction
Pre	Precision score
Rec	Recall score
RNA	Ribonucleic acids
SLiM	Short Linear Motif
SN	Sensitivity
SP	Specificity
STT	Single task transfer
topK	TopK Predictive Rate
VH	Virus-human protein-protein interaction training set
$\|V^b\|$	Number of bacteria proteins
$\|V^h\|$	Number of human proteins
$\|V^v\|$	Number of virus proteins
Y2H	Yeast-two hybrid
043570	Zika virus taxon ID
33761	HPV 18 virus taxon ID
5-FoldCV	5-fold cross-validation
644788	Influenza A virus taxon ID
697049	SARS-CoV-2 virus taxon ID

For/Dedicated to/To my family and my curiosity

Chapter 1

Introduction

1.1 Motivation

Recent technological advancements resulted in major breakthroughs in many research fields, especially Computer Science and Biology. In 2001, it took $2.7 billion USD and almost 15 years to complete the first human genome sequence as compared to only thousands and days nowadays. Genetics technologies enable assessments at various levels of granularity, scales, and at different time points. Our understanding of humans, animals, and other living forms has been improved to an unprecedented extent over the past years, with more and more data generated every day. Besides the whole genome sequence, public databases now store knowledge about the mutations found in genes, the disruptions in gene expressions, the protein expression in a particular tissue/cell line, or the detection of abnormality in genetic factors that affect a patient's condition, etc.

Yet human analytical and learning capabilities are restricted to a small amount of data. This poses opportunities for machine learning (ML) models which can effectively process big complex data. Due to their exceptional power in manipulating and utilizing a large volume of data, ML techniques have been applied to solve a tremendous number of challenging problems, including but not limited to those in biology. ML models can help prioritize wet-lab experiment candidates, reason about the underlying phenomena, or transform massive heterogeneous data into clinically actionable knowledge, etc. ML approaches have been applied to analyze biomedical data [90, 94, 172], identify diagnostic biomarkers [74, 98, 155, 160, 208], predict disease prognosis [22, 74, 85, 104, 121], detect disease subtype and select suitable therapy [64, 146, 151, 163, 212], prioritize and validate drug target [143, 240], or re-position drug [72, 199, 232].

Nevertheless, existing ML systems do not work effectively for unseen data [51–54]. The reasons for such phenomena are deep-rooted in some non-trivial challenges associated with existing biological datasets.

1.2 Challenges in analyzing biological datasets

Firstly, biological datasets usually suffer from the *data scarcity and bias* problems. Consider, for example, the protein-protein interaction (PPI) data which

is composed of several types of interactions between two proteins. Such interactions are usually analyzed to understand the molecular properties of certain diseases or search for potential drug targets. For such data, human intra and inter-species PPIs form the majority of all existing data in public databases [44, 204]. Similarly, most genome-wide gene expression and protein expression data come from human studies, and most studies only have up to 30 samples [165, 180, 210]. As another example, consider the miRNA-disease associations that encapsulate the pairwise connections between miRNAs and diseases. Such data can unveil a better understanding of diseases' pathology. Nevertheless, the majority of the currently known associations account for only a few well-studied diseases [52]. Models trained on such scarce and biased data tend to overfit, and their predictions cannot be generalized on new data. Moreover, due to limited amounts of training data, the potential of recent deep learning techniques cannot be fully utilized.

Secondly, biological datasets often encounter the *curse of dimensionality* problem [112]. That is to say, the dataset contains a huge number of features but has a very limited number of annotated samples. For example, most of the gene expression and protein expression datasets contain the expression information for thousands of genes but only have several to hundreds of samples [180]. The use of 'large feature space and small number of samples' to train ML models leads to unstable and non-generalizable machine learning models' performance [14]. As a consequence, when aiming for joint learning from multiple information sources, one cannot naively concatenate all the data, and more innovative methods are required.

Thirdly, biological datasets contain many *false-positives*. That is to say, the data contains many samples which are marked as positive but are actually negative. For example, considering the PPI data, it is estimated that the portion of false positives in public databases could be as high as 80% [81]. For miRNA-disease associations, according to the data deposited in the HMDD databases [88, 129], the number of false positives associated with a disease could be more than three-fold the number of true positives [52]. Low training data quality raises the concern regarding machine learning models' prediction capability.

1.3 Thesis scope and contributions

In this thesis, for simplicity, we refer to any disease, miRNA, protein-coding gene, protein, and other biological objects or molecules as biological entities. The thesis focuses on the development of ML models for two specific biological problems: miRNA-disease association and virus-human PPI prediction. The miRNA-disease association prediction is considered a binary classification problem in which, given an input pair miRNA-disease (m, d), predict the probability (in [0,1] range) that miRNA m is associated with disease d. Similarly, the virus-human PPI prediction is also treated as a binary classification problem in which given a pair of virus and human proteins (v, h), predict the probability (in [0,1] range) that the virus protein v interacts with the human protein h. The detailed problem definitions are given in chapters 3 and 7.

Being aware of the associated data challenges as described in Section 1.2, we identify the limitations of existing systems and propose our solutions.

1.3.1 Limitations of existing systems

We identified three types of issues associated with existing works in two biological problems as described in the following.

Degrading performance quality for new biological entities

We experimentally show that the data challenges often result in drastically downgraded prediction performance of existing models on unseen biological entities [51–54]. In chapters 5, 6, and 7, we give details regarding the problem and our proposed approaches.

Issues in the evaluation setup

We identify four main issues associated with many existing ML model evaluation setups. The first and most critical issue is the data leakage problem as described in detail in Chapter 4. Data leakage refers to the problem in which the testing data is observed during the model training process. Data leakage leads to over-estimation of models' performance and unfair comparison between models and is undesirable.

The second issue is the *unrealistic assumptions* on the number of negative samples in benchmarked datasets. Many existing systems assume the numbers of negative samples are equal to those of the positives. Nevertheless, for the PPI prediction problem, for example, most proteins only interact with a limited number of other proteins, which is much smaller than the number of all proteins ($\sim 20,000$) [142]. Similarly, while the number of known miRNAs is around $2,000$, the most well-studied diseases have at most several hundreds of associated miRNAs. Therefore, the assumption of balanced classes is unrealistic and is insufficient to quantify the systems' performance in realistic scenarios.

The third issue attributes to the evaluation strategies. Many existing systems focus only on the *transductive* testing setup in which the training data already contains some labeled data for the entities in the testing data. Nevertheless, they do not quantify how good their proposed models and others are in the inductive testing setup, where there exist completely new entities whose prior known associations/interactions are not observed in the training data. Yet our understanding of humans and the environment is still far from complete, and due to the limited time and resources, existing knowledge is often biased towards only some well-studied entities. Evaluating machine learning models only on the partly known entities neglect an important aspect of the system performance on unseen data.

The final issue is related to the use of improper evaluation metrics. Many existing works rely on the Area under the Receiver Operating Characteristic (AUC) or the Area under the Precision-Recall Curve (AUPR) scores as the main evaluation criteria. Nevertheless, such evaluation metrics can result

in misleading or overestimation in some cases [51, 235]. Such evaluation metrics would potentially lead to unfair and deceitful comparison among models.

Unrealistic case studies or use cases

Existing works for biological problems, for example, the miRNA-disease association prediction, often include case studies to showcase the system utilities in helping field experts to select potential candidates for wet-lab experiments. However, many existing systems' case studies often employ artificial subsets of the possible search space and mainly focus on well-studied biological entities. Such approaches are delineated from the real use cases since (i) they place unrealistic assumptions on the testing subsets and (ii) the number of well-studied entities is very limited while the number of unknown or little-known entities is overwhelming.

1.3.2 Our proposed solutions

To overcome the existing systems' limitations and the challenges associated with biological datasets, we here present our proposed solutions.

Joint learning models to improve prediction performance and data quality

We focus on ML systems for biological problems in which the labeled data is scarce. We aim to develop models that can effectively mitigate the data scarcity and bias problem. At the same time, such approaches should also be able to generate reliable predictions for new biological entities, i.e., the ones that have not been observed during training.

We notice that besides the given annotated data, there also exists various related biological knowledge. For example, for the miRNA-disease association prediction problem, apart from the limited miRNA-disease known associations, there also exist the data corresponding to the miRNA family, the disease ontology, as well as the miRNA-protein coding gene (PCG) and disease-PCG associations, etc. The miRNA-PCG and disease-PCG associations store biologically rich features that affect the miRNA-disease association probabilities. In contrast, the miRNA family and the disease ontology enclose the functional similarities/differences between miRNAs and diseases. Similarly, for the virus-human protein-protein interaction (PPI) prediction problem, besides the inter-species virus-human PPIs, we can retrieve the human PPIs and the abundant sources of unannotated protein sequences from public databases. The knowledge acquired from the human intra-species PPI network stores information related to human proteins' interaction/binding patterns. At the same time, the abundant source of protein sequences encapsulates the *language* of the proteins that is a rich information source for protein representation learning.

In short, we develop effective ML models that can jointly learn from heterogeneous biological information sources. Such a learning strategy is intuitive and is motivated by humans. When a person encounters a new problem, he or she tends to exploit past skills and experiences to solve it. From a computational perspective, those past skills and experiences are domain knowledge that is stored in the related data. Combining diverse information sources helps compensate for missing or unreliable information in each individual source. At the same time, multiple knowledgebases centering around common biological entities can offer global insights into the problem and, thus, help increase the overall reliability of the prediction.

Yet our employed data often exists in different formats and requires further pre-processing effort. In addition, because of the 'high dimension' nature of biological datasets, simple concatenation does not work, especially when the number of annotated samples is limited. Therefore, our focus is on ways to incorporate such related information sources without further increasing the data dimensionality. We concern with three main research questions:

- Which are the available sources of information that can be incorporated?
- How can we integrate such data?
- How can we control the quality and quantity of the added information?

To answer the above questions, we first identify the related information sources and their biological relevance. From that, we propose simple yet effective models that offer flexible ways to fuse various information sources at different stages of the model-building processes. It is essential that the model has to be simple because a complex learning architecture with a bulky parameter set would be prone to overfitting, given the limited training samples. Regarding the quantity and quality of the added side information, we employ various data filtering techniques, from a naive threshold-based approach to more complex methods based on expert domain knowledge.

In our joint learning frameworks, the integrated data sources can be utilized as sources to construct the feature space, as sources for statistical embedding learning, as the training data for the added side task(s) in a multitask learning framework, or as the supervised signals for the feature enrichment and filtering module. We employ the architectures proposed for language modeling, multitask learning, graph representation learning, and feature selection to construct our joint learning models. The diverse incorporated information sources are exploited to inform the learning module with the related domain knowledge or to guide the data preprocessing unit. The integrated side data not only helps *mitigate the data scarcity* and bias problem, but also claims their benefits in *controlling the quantity and quality* of the added information.

Our proposed models gain state-of-the-art performance in two biological problems, even for new or little-known biological entities. They also show great potential in *overcoming the high false positives problem in the annotated*

training data. More details regarding our joint learning approaches are presented in chapters 5, 6, and 7.

Consistent evaluation framework

We address the issues related to existing works evaluation setups (ref. Section 1.3.1) in three ways. **Firstly**, we develop a consistent evaluation framework with the modifications for state-of-the-art methods' training workflow as well as the implementation for various feature generation methods to *overcome the data leakage problem*. At the same time, our proposed joint learning approaches can also avoid such critical issue because they exploit the related knowledge sources to learn the corresponding entities' representations without resorting on any pre-defined feature generation strategy. **Secondly**, we propose replacing the widely adapted metrics with a *consensus evaluation metric* that can address the identified limitations to enable fair comparisons between models. **Lastly**, we introduce the use of an *inductive* testing setup to assess the methods' performance on new biological entities. At the same time, we create and release new datasets that consider the nature of biological data with realistic negative sample rates in various testing scenarios. Thus, enabling comprehensive comparisons among compared approaches. We believe such an evaluation framework, consensus metric, and new datasets will offer fair playgrounds and facilitate future research.

Model realistic use cases and support for field-experts' assessment and adoption

This section presents our solution regarding the issue related to unrealistic case studies as well as our effort in facilitating future research and applications.

Realistic use cases or case studies

We are the first to add in *realistic* case studies in which:

- We place no simplified assumptions on the potential candidate search space.
- We showcase the systems' utilities for completely new or little-known biological entities.
- We demonstrate the systems' applicability in differentiating between true positive and false positive samples, which is of great interest to field experts.

Chapters 5, 6, and 7 presents more details regarding our realistic case studies.

An easy-to-use application

One crucial factor that prevents the applicability of any ML model is usability. ML approaches that are scattered on the Internet without a detailed user guide or an easy-to-use application will not be widely adopted among the research communities. We tackle this problem by developing and releasing a

web application for the miRNA-disease association prediction that integrates all our system's generated prediction results and related domain information to facilitate field experts' assessments and adaptation. Chapter 6 gives details regarding our web application.

1.4 Summary of thesis contribution

In the following, we summarize our thesis contributions:

- We identify and experimentally analyze different types of issues associated with existing works.
- We develop a consistent evaluation framework with consensus evaluation metrics and new datasets to enable large-scale, fair, and comprehensive comparisons between models.
- We propose novel joint learning models that can flexibly and effectively integrate diverse information sources to improve ML systems' performance on two biological problems. The proposed learning frameworks acquire state-of-the-art prediction performance, even on new biological entities. Also, the systems show great potential in overcoming the high false positives problem in training data.
- Finally, we add realistic case studies and an easy-to-use web application to facilitate future research and adaptation.

1.5 Thesis layout

In Chapter 2, we start by introducing the related works regarding machine learning techniques for supervised classification problems that can exploit additional information sources besides the given data for the target learning task. We provide a brief overview of multimodal learning, semi-supervised learning, transfer learning, and multitask learning, as well as their similarities and differences with regard to our work. In addition, we summarize the general biological background associated with our studied biological problems in section 2.5.

Chapter 3 discusses the background knowledge specific to the miRNA-disease association prediction task. In section 3.1, we present an overview of the publicly available miRNA-disease association data and their limitations. Then in section 3.2, we state a formal problem definition followed by an introduction to various types of miRNA and disease similarities and features employed by existing works. Finally, in section 3.3, we briefly review the related works in miRNA-disease association prediction along with a detailed description of seven computational models and some of their variants that are employed in our experiments.

Chapter 4 encapsulates our first experimental work in miRNA-disease association prediction. We present a detailed analysis and discussion of existing systems' limitations. From that, we propose our solutions and recommendations.

Chapter 5 continues our work for the miRNA-disease association prediction with the proposal of a multitask learning framework that can exploit information from five different related knowledge sources and overcome most existing systems' issues. In addition to the new model development, we put forward the employment of the inductive testing setup with the curation and release of new datasets to evaluate benchmarked models on new miRNAs and new diseases that have not been observed during training. Moreover, we introduce a realistic case study corresponding to a disease with many false positives. We vary the false positive rate in the training data and evaluate how well the models can identify the true positives in section 5.4.1.

As the multitask model presented in chapter 5 still has some limitations concerning the quality and quantity of the integrated data, we present a solution to such issues in chapter 6. We propose a biological-driven message passing framework for miRNA-disease association prediction with a parameter-free mechanism to enrich and filter the integrated information sources in section 6.1. In addition, we add a realistic case study for a disease with scarce knowledge in section 6.5.1 and a survival analysis on publicly available miRNA survival and expression data in section 6.5.3. Finally, we add our support for end-user assessment and adoption by introducing a web application with all the related biological knowledge in section 6.6.

Chapter 7 is devoted to our work on the virus-human PPI prediction problem. We start with a brief introduction, the biological background, and the associated challenges in section 7.1. We then summarize the related works in section 7.2. Next, we present our proposed multitask learning framework that can exploit the knowledge from the abundant source of $\sim$24 million unannotated protein sequences and the knowledge acquired from the intra-species human protein-protein interaction network in section 7.3. Finally, we introduce a realistic case study with promising results for the SAR-Cov-2 human receptor prediction task in section 7.6.

Lastly, we give the conclusion and a discussion regarding the future work direction in Chapter 8.

1.6 List of publications

The core contributions of the thesis are presented in the following publications:

- The contributions in Chapter 4, which identify and analyze limitations of existing systems as well as present our proposed solutions and recommendations for the miRNA-disease association prediction, are published in:

  - **Thi Ngan Dong**, Megha Khosla, "*Towards a consistent evaluation of miRNA-disease association prediction models.*", 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 1835-1842, 2020.

- The contributions in Chapter 5, which proposes a multitask learning framework for miRNA-disease association prediction, along with the proposal of inductive testing setup, new datasets, and a new realistic case study for a disease with many false positives, are presented in:

  - **Thi Ngan Dong**, Stefanie Mücke, Megha Khosla, "*MuCoMiD: A multitask graph Convolutional Learning Framework for miRNA-Disease Association Prediction*", in IEEE/ACM Transactions on Computational Biology and Bioinformatics, doi: 10.1109/TCBB.2022.3176456, 2022.
  - **Thi Ngan Dong**, Megha Khosla, "*A multitask Convolutional Learning Framework for miRNA-Disease Association Prediction*", BIOKDD 2021.

- The contributions in Chapter 6, which proposes a message passing framework with multiple data integration for miRNA-disease association prediction, the new inductive testing sets for new disease evaluation, and the two new realistic case studies are presented in:

  - **Thi Ngan Dong**, Johanna Schrader, Stefanie Mücke, Megha Khosla, "*A Message Passing framework with Multiple data integration for miRNA-Disease association prediction*", Scientific Reports, volume 12, start page 16259, 2022.

- The contributions in Chapter 7, which proposes a multitask learning model for virus-human protein-protein interaction prediction with a realistic case study for SAR-Cov-2 human receptor prediction, are presented in:

  - **Thi Ngan Dong**, Graham Brogden, Gisa Gerold, Megha Khosla, "*A multitask transfer learning framework for the prediction of virus-human protein-protein interactions*", BMC Bioinformatics, volume 22, start page 572, 2021.
  - **Thi Ngan Dong**, Megha Khosla,, "*A multitask transfer learning framework for Novel virus-human protein interactions*", ICLR Workshop on AI for Public Health, 2021.

During the early stages of the Ph.D. studies, I also investigated machine learning feature selection methods with respect to the training and testing data complexity. Due to space limit, such an aspect is not touched in this thesis but is published in:

  - **Thi Ngan Dong**, Megha Khosla. "*Revisiting Feature Selection with Data Complexity for Biomedicine*." IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), pages 211-216, 2020.

From 06.2020 to 12.2020, I worked on patient clinical data retrieved from the Big Data for Cochlear Implant project. We tried to address various research questions centering around the patient cochlear implant outcome prediction. Due to space limit, such contributions are not encapsulated in this thesis but were presented in an L3S internal report:

- **Thi Ngan Dong**. "*Big Data for Cochlear Implant Project report*", L3S technical report, 2020.

Starting in early 2022, I join the COVID-19 project and have been working on patient gene expression data. We develop and release an end-to-end computational framework that encapsulates the variety of "gene selection" methods with different selection criteria and objectives. In addition, we create and release new datasets as well as propose new evaluation metrics for the purpose of fair and comprehensive comparison between different approaches. Such contributions are not included in this thesis but are presented in:

- **Thi Ngan Dong**, Megha Khosla, "*A consensus multi-perspective evaluation framework for feature selection over gene expression data*" in preparation.

In addition, the following publication was also completed over the course of this thesis:

- Tianqi Zhao, **Thi Ngan Dong**, Alan Hanjalic, Megha Khosla, "*Multi-label Node Classification On Graph-Structured Data*" submitted to the Learning on Graphs Conference (LoG) 2022.

Chapter 2

Background and Related work

The motivation for our work, as many other related works, is that collecting and cleaning annotated training biological data for a particular machine learning problem is often expensive, time-consuming, and even unrealistic in some scenarios. To overcome this issue, we exploit the presence of multiple types of information available pertaining to a problem. To that end, we develop effective joint learning models inspired by multimodal, transfer, and multitask learning techniques. In the following, we present the basics of these core learning techniques and highlight their similarities and differences from ours. Besides, we present the necessary biological background.

2.1 Multimodal learning

Modality refers to the way in which something happens or is experienced. A research problem is considered multimodal when it includes multiple such modalities. Figure 2.1 presents an example of a multimodal learning system in which the model learns the latent representation from the input images, text, and audio separately and then later concatenates them to form the input representation for the target prediction task.

In general, multimodal datasets contain data from different modalities, *observing common phenomena*, and are created with the targeted usage of complementary utilization toward learning a complex task. It is very common that an image caption contains information that is not conveyed in the image. Such information can be the metadata related to the capturing context, the author, the time and location, etc. Similarly, sometimes it is more straightforward to use an image to describe the information, which may not be obvious or hard to explain from the texts. For instance, using emotion icons or an X-ray or Magnetic Resonance brain Image (MRI) to present details regarding the patient's condition would be more intuitive than a piece of text.

Multimodal learning and our work both focus on jointly learning from multiple sources. Also, they both have to deal with the problems associated with the input data quality, for example, the varying level of noise and conflict between modalities, or the data scarcity, bias, and false positives in the added information sources. Nevertheless, our work differs from multimodal learning in two main points.

Firstly, the integrated data in multimodal learning applications contain several types of information related to the same input. It is obvious that any

single source could be employed to construct the feature space for the targeted learning problem. The patient clinical data or the diagnosis images alone can be utilized as the stand-alone input source for the patient-related classification task. Nonetheless, not all data sources in our integrated data pool can be employed as raw input for our targeted classification problem. Most of them can only be utilized as *side information sources* that encode complementary domain knowledge related to the entities in our training and testing data.

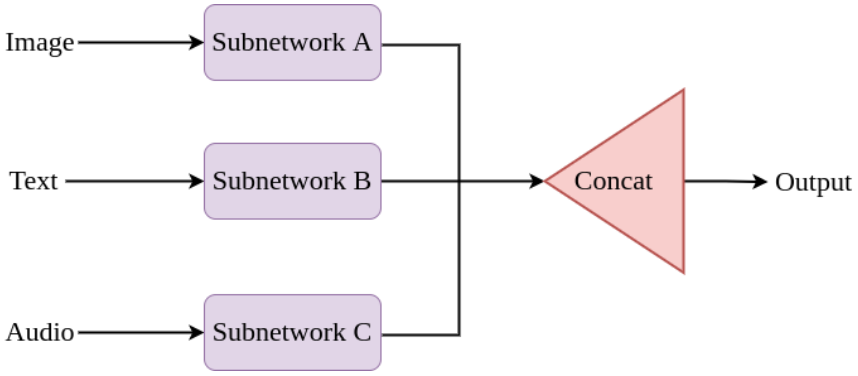


FIGURE 2.1: A multimodal learning system [110].

Secondly, from a technical perspective, muti-modal learning often involves the utilization of deep learning techniques that learn high-level embeddings from the multimodal data and then later combine them to construct a joint representation as input to a supervised prediction system [174]. In contrast, our work focuses on *simple models* with *flexible information integration strategies*. It is essential that the model has to be simple and requires a minimal set of parameters. Because with limited training data, a complex model with a large parameter set would easily be prone to overfitting.

2.2 Semi-supervised learning

Semi-supervised learning [23] is a machine learning methodology that lies between supervised and unsupervised learning. Semi-supervised approaches are often employed in the context where we can acquire massive unannotated data while only having very limited labeled samples. An illustration of a semi-supervised learning system is presented in Figure 2.2. The learning objective, in this case, is usually to utilize those unlabeled data to improve the target classification task. For example, we can retrieve millions of unannotated text documents from the Internet but only have hundreds to thousands of annotated documents for our text classification task. A semi-supervised approach, within this context, will train a self-supervised language modeling model on such a numerous pool of unannotated documents to learn statistical representations for text documents and then feed such representations as input to train a supervised text classification model on the limited labeled data.

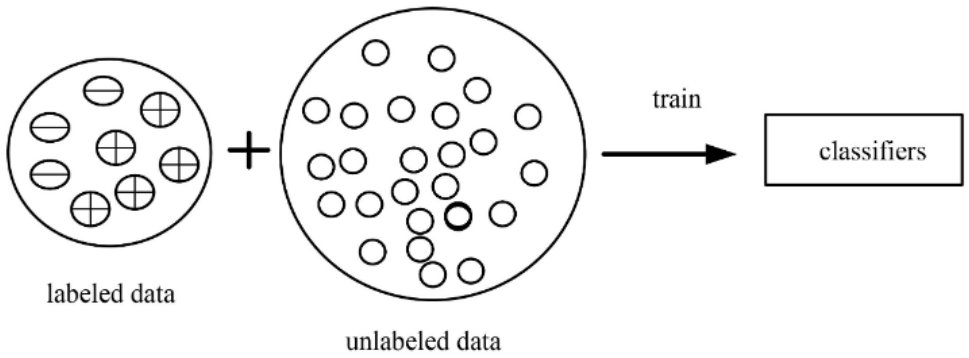


FIGURE 2.2: A semi-supervised learning system [245].

To some extent, semi-supervised learning is just one specific technique employed in our proposed models. Our joint learning approach can exploit unlabelled data, like the abundant source of unannotated protein sequences, to improve the target classification task, like virus-human protein interaction prediction. At the same time, we can flexibly utilize other techniques, like message passing on the protein-protein interaction network to enrich our feature representation or feature selection with the disease category information to filter redundant and noisy features.

2.3 Transfer learning

Transfer learning (TL) [211, 247] is a methodology that focuses on improving the performance of target machine learning models on target domains by transferring the knowledge contained in different but related source domains. An example of a transfer learning system is presented in figure 2.3. The given transfer learning model first trains a ML model for some supervised prediction task on the large annotated source data and then fine-tunes that model using the small number of annotated samples corresponding to the target prediction problem (which is different from that of the source). Transfer learning methods can be categorized into homogeneous and heterogeneous approaches depending on the input representation of the source and target domains.

Homogeneous transfer learning [247] assumes that both source and target domains share the same feature space and only differ in the marginal feature distribution. To some extent, homogeneous transfer learning has connections with semi-supervised learning, which transfers the knowledge stored in an abundant source of unannotated data to improve the input representation for the target prediction task. Nevertheless, homogeneous transfer learning covers a broader application context. For example, assume that we have a rich source of annotated image samples for the car recognition problem but only a limited number of labeled images for the truck recognition task. A transfer learning approach, in this case, could be used to train a supervised classification model to recognize whether the input image contains a car and then fine-tune that model for the truck recognition problem.

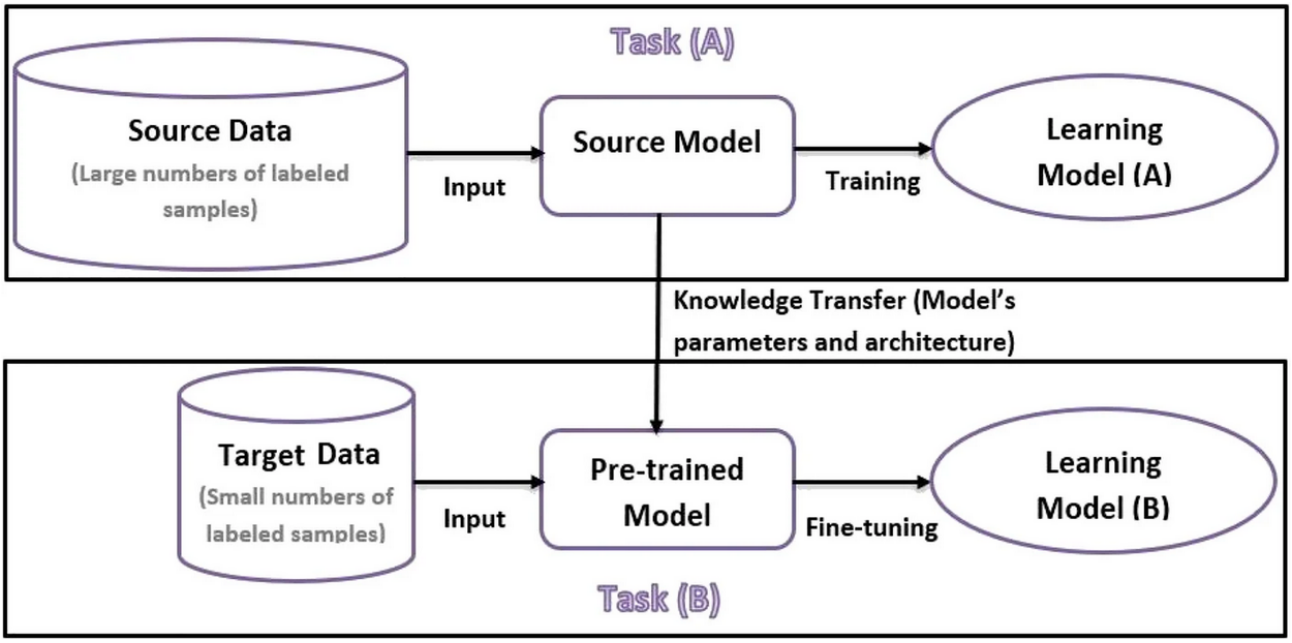


FIGURE 2.3: A transfer learning approach [59].

Heterogeneous transfer learning [247] aims at transferring knowledge while the domains have different feature spaces. In other words, the related knowledge is available in the source data, but it is represented in a different way other than that of the target. Let us take an example. Given that there are very limited labeled image samples but much more abundant annotated Web documents that contain images. One could employ a heterogeneous transfer learning approach to transfer the semantic knowledge stored in those text documents to improve the target image classification task.

In general, our approach resembles existing works in transfer learning in many points. Nevertheless, regarding the data sources, our approach is broader. We can exploit the knowledge from completely different information sources, while most existing transfer learning approaches cannot. For instance, our target problem can be the miRNA-disease association prediction when our learning sources can be the protein-protein interaction data.

2.4 Multitask learning

Multitask learning [21] is an ML paradigm that aims at performance improvement through simultaneously learning from multiple related tasks. Transfer learning can be considered as an asymmetric modification of multitask learning where there is an explicit distinction between the source and target tasks. An example of a multitask system is presented in figure 2.4 in which the multitask model tries to generate predictions for four tasks simultaneously from the same input. Multitask learning is particularly appropriate when there are a large number of related tasks and/or each task has only limited labeled samples.

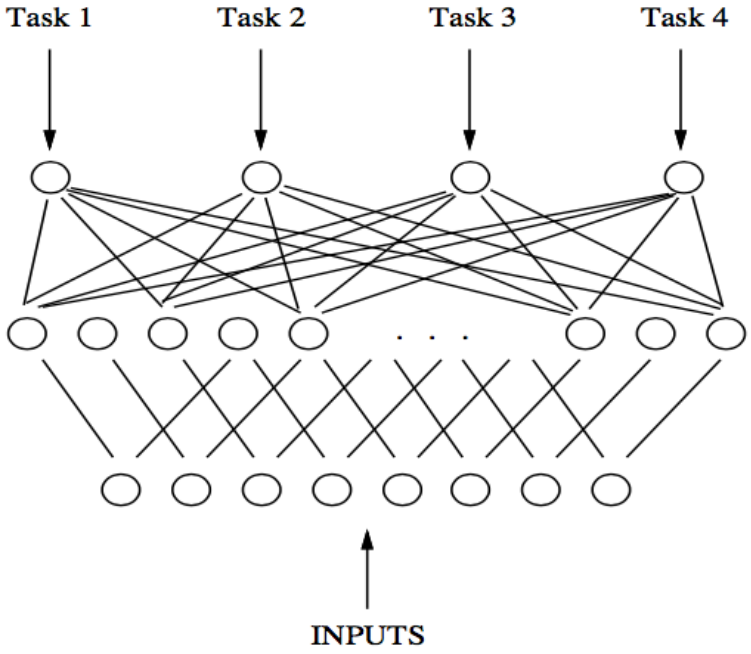


FIGURE 2.4: An example of a multitask system [21].

Both multitask learning and our approach aim at learning from multiple sources to improve a supervised prediction's performance. Nevertheless, to some extent, our work is much more flexible. For example, we can simultaneously learn to predict both virus-human PPI and human-human PPI. However, our proposed system can be a multi-stage model in which we asynchronously do feature selection according to the disease category and shallow node embedding learning from a heterogeneous graph to improve our target miRNA-disease association prediction task.

2.5 Biological background

2.5.1 DNA and gene

The human genome is encoded in all cells as DNA. *DNA*, or deoxyribonucleic acid, refers to a double helix strand of nucleotides that stores the genetic materials of human and almost all other organisms. Nearly every cell in the same organism has the same DNA and most human DNA sequences only differ by less than one percent. *Genes* are functional units of the DNA. Protein coding genes (PCGs) are the genes that encapsulate the instructions for protein generation. Other genes store the genetic information for non-coding RNAs. It is estimated that the human genome contains approximately 20,000 PCGs [187].

2.5.2 RNA and transcription

*RNA*s or Ribonucleic acids are chains of nucleotides that share similar structures to the DNA. However, unlike DNA, RNAs are mostly single strands.

Gene expression is the process of unzipping parts of the DNA into either RNA molecules that code for proteins (messenger RNAs, mRNAs, or transcripts) or non-coding RNA molecules that serve other functions. The conversion process from PCGs to mRNAs is called *transcription*. Though most human cells have the same DNA, the dissimilarity in gene expression enables them to have non-identical functions and structures. Different genes will have different expression levels under different conditions. Also, the expression of one gene can affect the expression of some other genes. This phenomenon is often referred to as *genetic interaction* or *gene functional interaction*.

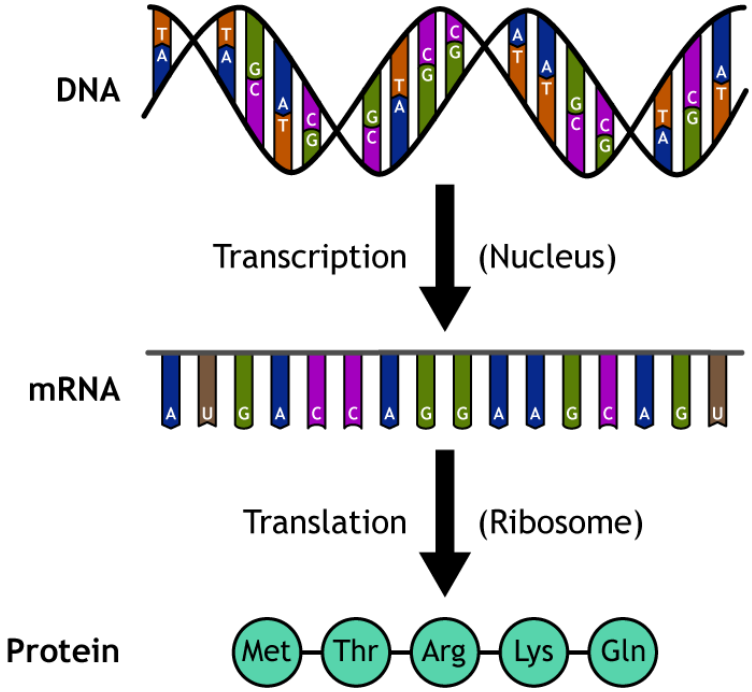


FIGURE 2.5: The relationship between DNA, RNA, protein, transcription and translation processes [83].

2.5.3 Protein and translation

Translation is the process of synthesizing proteins or sequences of amino acids from the genetic code stored in mRNAs. *Protein*s are large, complex molecules responsible for essential biological functions inside living organisms. Proteins rarely act alone but interact with others to carry out biological functions. *Protein expression* is defined as the complex mechanism in which proteins are synthesized, modified, and regulated inside living organisms. Disruptions in proteins' expressions are directly associated with various disease conditions [153]. Figure 2.5 presents an illustration of the relationship between DNA, RNA, protein, and transcription and translation processes.

2.5.4 Disease and miRNA

Micro RNAs or *miRNAs* is a highly conserved class of non-coding RNAs with a length of approximately 22 nucleotides. miRNAs fulfill their diverse functions by regulating the gene expression of PCGs after transcription. The transcribed mRNAs can be directly bound by miRNAs, which leads to cleavage or destabilization of the mRNAs and represses the translation into proteins [18]. Figure 2.6 illustrates miRNA regulation mechanism.

The binding between the miRNAs and their target mRNAs is facilitated by complementary base pairing between the so-called seed region of the miRNAs and the matching sequence in the mRNAs found most often in the 3'UTR [173]. Each miRNA can have hundreds of target mRNAs. Also, each mRNA can be regulated by more than one miRNA. Though this complicated regulatory network is not yet fully understood, it is estimated that about one-third of all PCGs is regulated by at least one miRNA [184]. These ubiquitous regulatory functions are also responsible for the multitude of cell processes influenced by miRNAs: cell development, maturation, differentiation, and apoptosis as well as cell signaling, cellular interactions, and homeostasis [65, 108, 147, 183]. Consequently, the mutation of miRNAs or changes in their expression can have diverse consequences that can be hard to predict. Recent studies indicate that miRNAs could serve as potential biomarkers in certain diseases such as cancers or immune-related diseases [97, 105, 133, 179, 189, 207, 242].

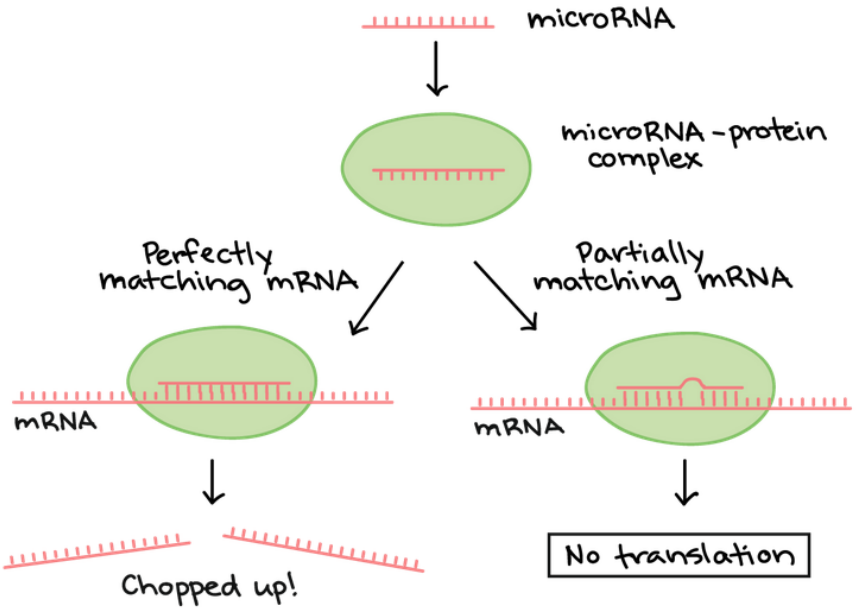


FIGURE 2.6: miRNA regulates protein translation [1]

A *miRNA family* is the group of miRNAs that share a common ancestor in the phylogenetic tree. MiRNAs that belong to the same family usually have highly similar sequence secondary structures and tend to execute similar biological functions [102]. Similar miRNAs would tend to participate in the mechanisms of similar diseases. Figure 2.7 presents the sequences of some miRNAs in the miR-9 family [135].

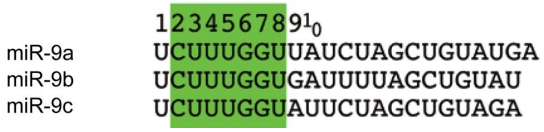


FIGURE 2.7: Some miRNAs in the miR-9 family and their sequences [135].

The *disease ontology* [190] represents the disease etiology classes. Each disease can have multiple subclasses and, at the same time, can be the subclass of one to several other diseases. Similar diseases can be expected to associate with similar miRNAs. Figure 2.8 gives an illustration of the disease ontology [52].

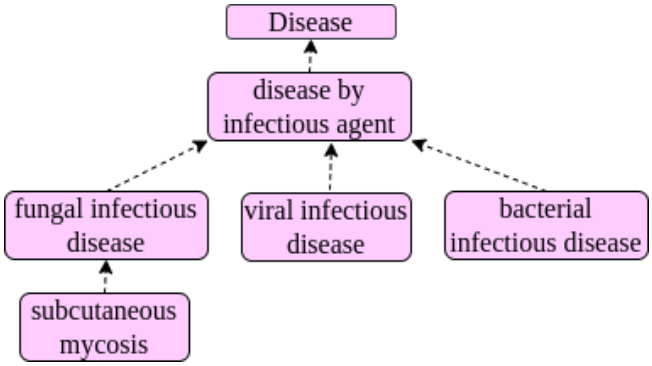


FIGURE 2.8: The disease ontology illustration [52].

2.5.5 Virus and infectious cycle

Viruses are the smallest known form of life that can only replicate inside living organisms (or hosts). The virus infection process, which is referred to as the *infectious cycle*, can be divided into different stages involving many protein-protein interactions (PPIs) between the virus and its host. These interactions range from the initial binding of viral coat proteins to the host membrane receptor [196], to the uncoating of the virus genome [39], hijacking of the host transcription machinery [213], and then assembling and release of the new viruses [39, 57]. Figure 2.9 presents an illustration of the infectious cycle. Understanding the PPI between a particular virus and its host plays a crucial role in unveiling the underlying mechanism of virus infection and pathogenesis.

Recent studies indicate that viruses' capsid protein sequences show little to no conservation. They are structurally dynamic such that they cannot be easily detected by common sequence-structure comparison [175]. Also, two viral proteins with entirely different sequences can share similar functions and interact with a similar set of human proteins.

2.5.6 Pathway enrichment analysis

A *biological pathway* is a series of actions/interactions among molecules (like genes, proteins) that can control cell chemistry, transmit signals, or regulate

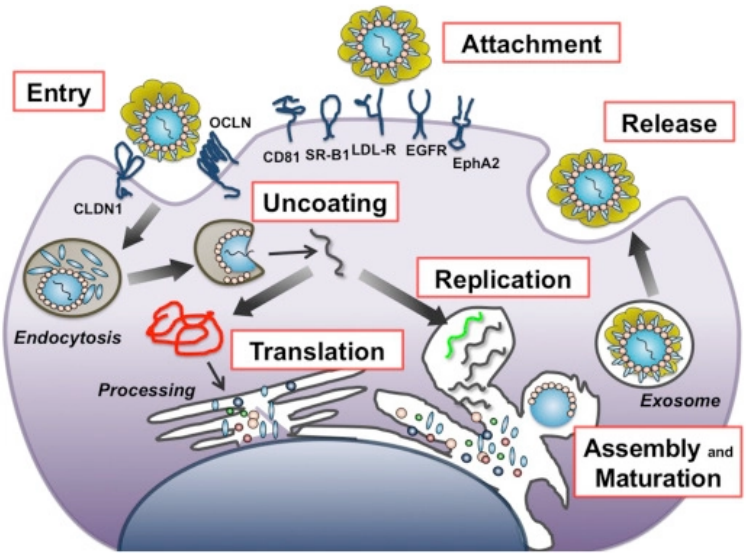


FIGURE 2.9: The virus infection cycle [57].

gene expression and, thus, can subtly determine how a person responds to the world. High-throughput experiments often yield large sets of genes/proteins that are nearly impossible to interpret manually and draw insights. *Pathway enrichment analysis* that aims at identifying subgroups that are significantly represented or under-represented in a larger input set of genes or proteins, in such a case, can provide field experts clues about the factors that drive the biological conditions. The analysis often involves utilizing statistical methods over an extensive knowledgebase of known pathways such as the Reactome pathway knowledgebase [61, 62] and returns a list of significantly enriched pathways and their statistical significance.

2.5.7 Functional enrichment analysis

Each gene is believed to have some pre-defined functions. Large knowledgebases like the Gene Ontology [20] are community efforts to organize and standardize our understanding so far on gene functions. *Functional enrichment analysis* is a statistical method built upon such knowledgebases to identify the functional properties that are significantly over or under-represented in a group of genes. From that, field experts draw insight into the results obtained from high-throughput experiments. Figure 2.10 provides an example of the pathway and functional enrichment analysis results.

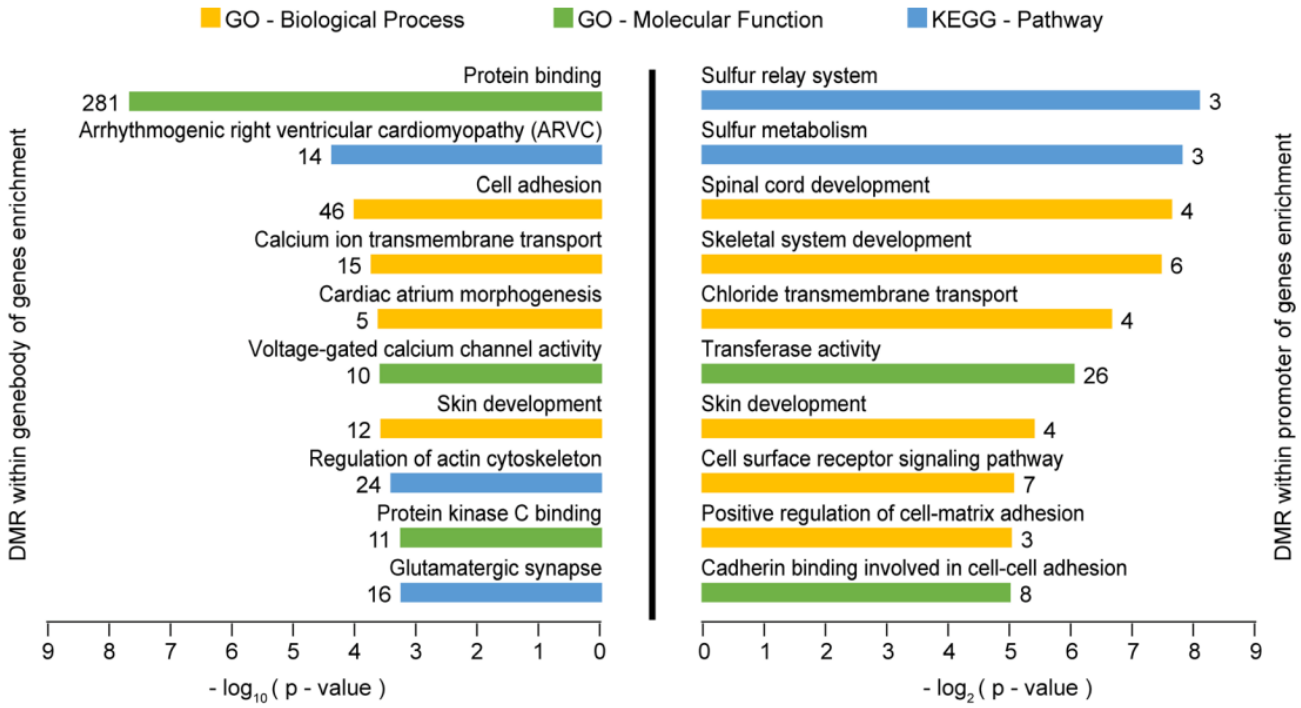


FIGURE 2.10: Pathway and functional enrichment analysis results [224].

Chapter 3

Predicting miRNA-disease association

MiRNA-disease association prediction has promising applications in drug development as well as disease diagnosis and treatment. In the past decade, hundreds of computational approaches have been proposed. Nevertheless, some crucial limitations still prevent existing works from generating reliable predictions, especially for new miRNAs and new diseases. This chapter summarizes the motivation, background, and related works corresponding to machine learning models on the miRNA-disease association prediction problem.

3.1 Introduction

One pivotal element for the success of a drug development process is drug target selection. Due to the high number of failures in the protein-target drug trial period, the target in drug selection has started shifting to RNAs. MicroRNAs (or miRNAs) are becoming promising drug targets as they can involve in many biological processes (from cell duplication, propagation, cell death, etc., to more complicated mechanisms like virus-host interaction) [25, 26, 28]. Identifying potential associations between miRNA and disease would help in clinical diagnosis, treatment, and drug development. Since wet-lab experiments are expensive and time-consuming, recent years have observed an upsurge in the number of proposed machine learning-based computational approaches.

Regarding the data, the curated HMDD v2.0 [129] and HMDD v3.0 [88] databases are the most complete and widely used information sources among existing works. As is typical to many biological applications, a major challenge for building generalizable and eventually well-performing models for the miRNA-disease association prediction problem is **data scarcity**. For example, the total number of miRNAs and diseases in the standard HMDD v2.0 [129] database are 578 and 383, respectively, with a total of 6,447 associations (without any preprocessing and removal of duplicates). Accordingly, the number of known associations is at most 3% the number of possible miRNA-disease pairs. Such a small and sparse dataset prohibits the utility of flexible and expressive modern representation learning techniques.

Even worse, the known association set is biased towards some well-studied diseases (∼10% of the most well-studied diseases account for ∼60% of the known associations, and ∼20% of the most well-studied diseases account for ∼80% of the known associations). Figure 3.1 presents a disease frequency graph in which the y-axis refers to the number of diseases while the x-axis denotes their corresponding number of known associations found in the HMDD databases. It can be seen that most diseases have less than 20 known associations, while some of them can have hundreds.

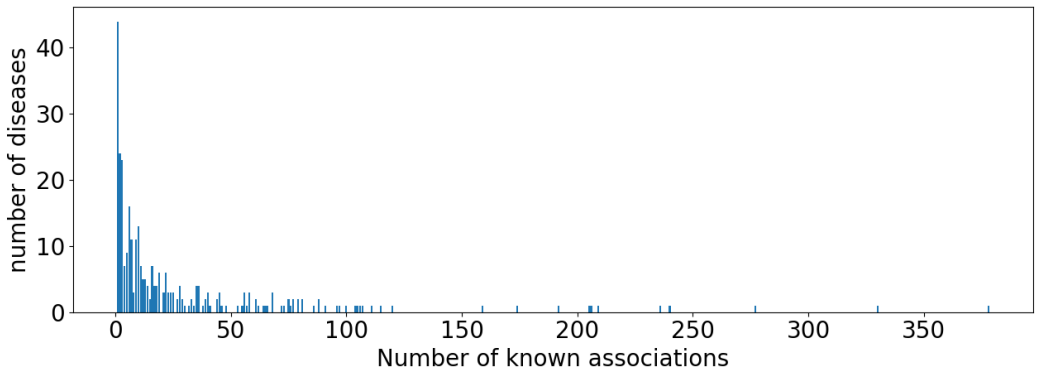


FIGURE 3.1: The disease frequency bar graph.

In addition, known miRNA-disease association data contains many false-positives. In one of our case study, we discover that the number of false positives corresponding to a disease could be more than three folds that of the true positives [52]. These data challenges often leads to *biased* and *non-generalizable* models.

3.2 Background

3.2.1 Problem definition

Given the set of known miRNA-disease associations, we treat the miRNA-disease association prediction as a binary classification problem where the label for an input pair node (m, d) is 1 if there is a known association between them and 0 otherwise. The machine learning model task, in this context, is to predict a probability in the [0,1] range, indicating how likely there exists an association between a given miRNA-disease input pair.

3.2.2 Similarity metrics

Due to the lack of information, most existing works often utilize some similarity measured among miRNAs or diseases as features. In this section, we present a brief overview of the most commonly used miRNA and disease similarities.

Disease Semantic Similarity (DS)

Existing works often acquires the disease ontology (as described in Section 2.5.4, Chapter 2) from the MeSH [15] disease descriptor database. For brevity, we refer to this as the MeSH ontology. A disease directed acyclic graph (DAG) is a directed graph constructed from a disease ontology (e.g., the MeSH ontology) where the nodes represent diseases, and there is a directed link between node d_i and node d_j if disease d_i 'is-a' specific instance of disease d_j or disease d_j is the parent of disease d_i. The disease semantic similarity[219] measures the similarity of two diseases based on their relative positions on the disease DAG. The contribution of disease d_i to the semantic value of disease d_j is then defined as:

$$\begin{cases} C(d_i, d_i) = 1 \\ C(d_i, d_j) = \max\{\Delta \cdot C(d_i, d_k) | d_k \in \text{children of } d_j\}, \text{ if } d_j \neq d_i \end{cases} \tag{3.1}$$

where Δ is the decaying factor and is set to 0.5 as in [219]. Then the semantic value for disease d_i is calculated as $S(d_i) = \Sigma_{d_j \in \mathcal{D}_i} C(d_i, d_j)$, where $\mathcal{D}_i$ is the set of d_i's ancestors.

Finally, the **semantic similarity** between two diseases d_i and d_j is given by:

$$S(d_i, d_j) = \frac{\sum_{d \in \mathcal{D}_i \cap \mathcal{D}_j} (C(d_i, d) + C(d_j, d))}{S(d_i) + S(d_j)} \tag{3.2}$$

where $\mathcal{D}_i$ and $\mathcal{D}_j$ are the set of ancestors for diseases d_i and d_j, respectively. However, Xuan et al. [233] argue that the more a particular disease appears as an ancestor of other diseases, the less specific it is for a particular disease. Given the likelihood $p(d)$ of a disease node d appearing as ancestors of all other diseases, Xuan et al. [233] quantify the information content of d, $IC(d)$, as the negative log of the likelihood, i.e. $IC(d) = -\log(p(d))$. The phenotype similarity score between two diseases is then computed as:

$$S'(d_i, d_j) = \frac{\sum_{d \in \mathcal{D}_i \cap \mathcal{D}_j} (IC(d) + IC(d))}{\sum_{d \in \mathcal{D}_i} IC(d) + \sum_{d \in \mathcal{D}_j} IC(d)} \tag{3.3}$$

The updated similarity score between two diseases is then calculated as an average of their semantic similarity and phenotype similarity scores:

$$\mathbf{DS}(d_i, d_j) = \frac{S(d_i, d_j) + S'(d_i, d_j)}{2} \tag{3.4}$$

For simplicity, from now on, without explicitly giving the formula, by saying **disease semantic similarity**, we mean the disease semantic similarity given in Equation (3.4).

MiRNA functional similarity (MF)

The miRNA functional similarity (**MF**) [219] relies on the assumption that two miRNAs are more similar if they associate with similar diseases. **MF** is

estimated based on the similarities among associated disease sets. In particular, first, the similarity between disease d_i and a group of disease $\mathcal{D}$ is computed as the maximum value of the similarities between d_i and any member disease $d_j \in \mathcal{D}$:

$$DS(d_i, \mathcal{D}) = \max_{d_j \in \mathcal{D}}(DS(d_i, d_j))$$

Given that $\mathcal{D}_i$ denote the set of diseases associated with miRNA m_i, the functional similarity between two miRNAs m_i, m_j is then calculated as:

$$\mathbf{MF}(m_i, m_j) = \frac{\sum_{d_i \in \mathcal{D}_i} DS(d_i, \mathcal{D}_j) + \sum_{d_j \in \mathcal{D}_j} DS(d_j, \mathcal{D}_i)}{|\mathcal{D}_i| + |\mathcal{D}_j|}. \tag{3.5}$$

Gaussian Interaction Profile (GIP) Kernel Similarity (MG and DG)

Van et al. [116] first construct the miRNA-disease association matrix $\mathbf{A}$ from the set of known associations. Let n_m and n_d be the number of miRNAs and diseases. Then $\mathbf{A}$ is a 2D matrix of size $n_m \times n_d$. Let $\mathbf{A}_{ij}$ denotes the value of $\mathbf{A}$ at row ith and column jth. Then $\mathbf{A}_{ij} = 1$ if miRNA ith associates with disease jth and $\mathbf{A}_{ij} = 0$ otherwise.

Given the miRNA-disease association matrix $\mathbf{A}$, the GIP kernel similarity [116] between two miRNAs is defined as:

$$\mathbf{MG}(m_i, m_j) = \exp(-\lambda_m ||IP(m_i) - IP(m_j||^2), \tag{3.6}$$

where $IP(m_i)$ and $IP(m_j)$ correspond to the ith and jth rows of matrix $\mathbf{A}$, respectively. The parameter λ_m is calculated as:

$$\lambda_m = \frac{n_m}{\Sigma_{i=1}^{n_m} ||IP(m_i)||^2}$$

Similarly, the GIP similarity between two diseases is calculated similarly as:

$$\mathbf{DG}(d_i, d_j) = \exp(-\lambda_d ||IP(d_i) - IP(d_j||^2), \tag{3.7}$$

where $IP(d_i)$ and $IP(d_j)$ correspond to the ith and jth columns of matrix $\mathbf{A}$, accordingly. The parameter λ_d is calculated as:

$$\lambda_d = \frac{n_d}{\Sigma_{i=1}^{n_d} ||IP(d_i)||^2}$$

MiRNA target similarity (MP)

Xiao et al. [230] first retrieve the gene functional interaction network from HumanNet [89] database. Each connection between two genes p_i and p_j is associated with a confidence score c_{ij} indicating how likely the connection is 'real'. If there is no functional linkage between the two genes, then $c_{ij} = 0$.

The similarity score between two genes p_i and p_j is then defined as:

$$SP(p_i, p_j) = \begin{cases} 1, \text{ if } p_i = p_j \\ c_{ij}, \text{ otherwise} \end{cases}.$$

Then the similarity between a gene p_i and a group of genes $\mathcal{P}$ is then defined as:

$$SP(p_i, \mathcal{P}) = \max_{p_j \in \mathcal{P}}(SP(p_i, p_j))$$

Let $\mathcal{P}_i$ denote the set of genes associated with miRNA m_i. Then a target similarity [230] between two miRNAs m_i and m_j is defined as:

$$\mathbf{MP}(m_i, m_j) = \frac{\sum_{p_i \in \mathcal{P}_i} S(p_i, \mathcal{P}_j) + \sum_{p_j \in \mathcal{P}_j} S(p_j, \mathcal{P}_i)}{|\mathcal{P}_i| + |\mathcal{P}_j|}. \tag{3.8}$$

Disease target similarity (DP)

The disease target similarity [205] is calculated according to the equation 3.8. Let $\mathcal{P}_i$ denote the set of genes associated with disease d_i. The disease target similarity score between two disease d_i and d_j is calculated as:

$$\mathbf{DP}(d_i, d_j) = \frac{\sum_{p_i \in \mathcal{P}_i} S(p_i, \mathcal{P}_j) + \sum_{p_j \in \mathcal{P}_j} S(p_j, \mathcal{P}_i)}{|\mathcal{P}_i| + |\mathcal{P}_j|}.$$

MiRNA sequence geometric similarity (MS)

Each miRNA can be represented as a sequence of nucleotides. Figure 2.7 presents some example of miRNA sequences. Li et al. [126] first map each miRNA sequence to a 2-dimensional (2D) representation, where each nucleotide in its sequence is converted to a 2-dimensional vector. Let nt_i denote the nucleotide at the ith position. The 2D representation x_i of nt_i is computed recursively as: $x_i = x_{i-1}\delta(x_{i-1} - \lambda_i)$ such that $x_0 = (0.5, 0.5)$ and the decaying factor $\delta = -0.5$. The parameter λ_i is set as follows

$$\lambda_i = \begin{cases} (0,0), \text{ if } nt_i = \text{A} \\ (0,1), \text{if } nt_i = \text{C} \\ (1,1), \text{if } nt_i = \text{G} \\ (1,0), \text{ if } nt_i = \text{U} \end{cases}.$$

The sequence geometric similarity [126] **MS** of two miRNAs m_i and m_j is then defined as the Euclidean distance between their two corresponding vectors in the above-defined 2D space.

MiRNA sequence alignment similarity (MA)

The miRNA sequence alignment similarity [205] between two miRNAs m_i and m_j is calculated according to the Needleman-Wunsch algorithm [157]

which employs dynamic programming to measure how similar two neucleotide sequences are. In particular, the algorithm first defines a score for each match, mismatch or match with a gap state between two individual neucleotides. Then it divides the original problem into smaller problems on smaller sub-sequences, assigns a score for each possible alignment, and then return alignments that result in the best scores.

Let NW store the alignment scores calculated by the Needleman-Wunsch algorithm for all pairs of miRNAs. The final miNRA sequence alignment similarity matrix $\mathbf{MA}$ is calculated such that:

$$\mathbf{MA}(m_i, m_j) = \begin{cases} 1, \text{ if } m_i = m_j \\ \frac{NW(m_i,m_j) - \min(NW)}{\max(NW) - \min(NW)}, \text{ if } m_i \neq m_j \end{cases}$$

3.2.3 Other types of input features

MiRNA family feature (MO)

MiRNAs belonging to the same family usually share similar secondary structures and have similar biological functions [102]. One miRNA family can contain many miRNAs, while one miRNA might not have the family information available. For those that do not belong to any known families, we assume it belongs to a new family whose name is its own. We model each miRNA family feature as the one-hot encoding of its family. In particular, let n_f be the number of all miRNA families. Then the miRNA family feature $\mathbf{MO}_i$ of a particular miRNA m_i is a vector of n_f values where a value $\mathbf{MO}_{ij}$ at position jth is set to 1 if m_i belongs to the jth family and 0 otherwise.

MiRNA target gene feature (MT)

As discussed in section 2.5.4, one miRNA can regulate the expression of hundreds of PCGs. The curated miRNA-PCG associations are usually retrieved from public databases like miRTarBase [86]. For each miRNA-PCG association, such databases calculate a confidence score based on multiple weighting criteria like the experimental method, the original data sources, etc., to justify how likely the association is 'real'. If the confidence scores are not in the [0,1] range, they are scaled accordingly.

Each miRNA target gene feature is then represented as a one-hot encoding or a weighted vector of its PCG associations. In particular, let n_p be the number of PCGs. Then the miRNA target gene feature $\mathbf{MT}_i$ of a miRNA m_i is defined as a vector of n_p values where a value $\mathbf{MT}_{ij}$ at position jth is set to be (i) 1 if there is a known association between miRNA m_i and the jth PCG and 0 otherwise (for the one hot encoding representation) or (ii) the association confidence score (in [0,1] range) between m_i and the jth PCG and 0 if such association record does not exist (for the weighted representation).

Disease associated gene feature (DT)

Disruptions in gene expression affect protein expression. Disruptions in protein expression might lead to diseases. Therefore, changes in gene expression are believed to associate with diseases. The disease-PCG associations can be retrieved from public databases like DisGeNET [168]. Such databases usually curate information from multiple sources under various experimental methods and conditions. For each association, a confidence score is often calculated to measure how likely such an association is real. If such a confidence score is not in the [0,1] range, they are scaled accordingly.

Each disease-associated gene feature can be modeled as the one-hot encoding or as a weighted representation of its associated PCGs. In particular, let n_p be the number of PCGs. Then the disease target gene feature $\mathbf{DT}_i$ of a disease d_i is defined as a vector of n_p values where a value $\mathbf{DT}_{ij}$ at position jth is set to be (i) 1 if there is a known association between disease d_i and the jth PCG and 0 otherwise (for the one hot encoding representation) or (ii) the association confidence score (in [0,1] range) between d_i and the jth PCG and 0 if such an association does not exist (for the weighted representation).

3.3 Related work

Hundreds of computational models have been proposed in the past decades for the miRNA-disease association prediction problem. Each approach often utilizes different architecture, input, experimental setups, and benchmarked datasets. It is impossible to discuss every approach in detail. This section presents a brief overview of existing works followed by a detailed discussion of seven recently proposed models and some of their variants included in our experiments.

3.3.1 Overview

Regarding the input data, existing works can be grouped into: those that rely on pre-calculated similarities, those that automatically learn miRNA and disease representations at run time, and the hybrid techniques.

Similarity-based methods

Similarity-based methods rely on pre-calculated similarities to construct their feature space. The similarity matrices can be directly used as input features or indirectly to build the input graph(s) for feature learning. Regarding the number of similarity metrics employed, similarity-based techniques can be further divided into single and integrated models. Single similarity-based approaches only employ two pre-calculated similarities for miRNA and disease (one for each). The model architecture can vary from a simple ranking model [219] to more complex systems that involve modern learning techniques like variational auto-encoders [49] or neural matrix completion [126]. Some additional examples of such methods include: the model from Chen

et al. [38], RWRMDA [30], NetCBI [27], RLSMDA [32], IMCMDA [37], EPMDA [55], and GCSENET [132].

Integrated models combine multiple such similarities together. The integration step can happen before or during the model training process. For those that happened before, a fixed formula is often employed to calculate a weighted combination of the multiple input similarities. Then the learning and prediction strategies are usually similar to those of the single similarity-based methods. For example, DBNMDA [34] combines two miRNA and two disease similarities to construct the representation for miRNAs and diseases. Similarly, NNMDA [239] proposes a weighted mechanism to derive the intra-connections among miRNAs and diseases from five miRNA and two disease similarity metrics. Some other examples for this type of methods include: HGIMDA [35], the model by Wei et al. [117], MDA-SKF [95], EDTMDA [33], LMTRDA [220], MSFSP [243], SCMFMDA [127], NCMCMDA [31], and SAEMDA [216].

Other similarity integration-based methods learn separate hidden representations from different similarity measures, then combine them later. DBMDA [244] and MMGCN [205] are two examples of such methods. The former simply concatenates the two hidden representations, while the latter relies on a multichannel attention mechanism to accomplish such a task.

Feature-learning-based techniques

The feature-learning-based techniques do not rely on pre-calculated similarities but automatically learn the miRNA and disease representations at run time. DIMIG 2.0 [162] and the model from Ji et al. [93] are two representatives for this type of methods. Both models employ a similar technique which integrates the information from multiple information sources to build large heterogeneous networks and then utilize graph representation learning techniques over the constructed networks to learn miRNA and disease representations.

Hybrid approaches

The hybrid approaches are combinations of similarity and feature-learning-based methods. NEMII [75] is one recent work that falls into this category. NEMII first construct a bipartite network from the known association data. The model then utilizes a structural deep network embedding (SDNE) technique to learn the structural embedding for miRNA and disease. The final miRNA-disease representation is formulated as the concatenation of their structural embedding, the miRNA family feature, and the disease semantic similarity (as described in section 3.2).

In the remaining subsections, we present details regarding seven recently proposed methods and some of their variants included in our experiments.

3.3.2 EPMDA [55].

EPMDA is a two stages method that consists of two disjoint modules, one for feature extraction and one for classification. Let n_m and n_d denote the number of miRNAs and diseases, respectively. The feature extraction module first constructs a heterogeneous network $\mathcal{G}$ from the set of known miRNA-disease associations (encoded by the association matrix $\mathbf{A} \in \mathbb{R}^{n_m \times n_d}$), the miRNA GIP similarity (stored in matrix $\mathbf{MG} \in \mathbb{R}^{n_m \times n_m}$), and the disease GIP similarity (stored in matrix $\mathbf{DG} \in \mathbb{R}^{n_d \times n_d}$). The weighted adjacency matrix ($\hat{\mathbf{A}} \in \mathbb{R}^{n_m+n_d \times n_m+n_d}$) of $\mathcal{G}$ is then defined as:

$$\hat{\mathbf{A}} = \begin{bmatrix} \mathbf{MG} & \mathbf{A} \\ \mathbf{A}^\top & \mathbf{DG} \end{bmatrix}$$

The feature extraction module then calculates a feature vector for each potential edge or link that connects any two points in $\mathcal{G}$. Let ℓ be an input parameter representing the maximum circle length. Then the feature vector of a particular input pair (m, d) is represented as a vector of ℓ values. Each feature value f_i corresponds to the disturbance level resulting from the addition or deletion of that particular edge to the set of cycles (in which it participates) of length i in $\mathcal{G}$. The higher the disturbance, the more important that particular edge is.

EPMDA's classification module is a 5-layer Multilayer Perceptron Regression model. For each miRNA-disease input pair (m, d), it takes the feature vector extracted from the previous step as input and outputs a probability in [0,1] range indicating how likely miRNA m is associated with disease d. To better avoid overfitting, EPMDA's classification model is also trained with **L2** or Ridge regularization.

Variants of EPMDA. We investigated a variant of EPMDA where we replaced the MLP regressor with a linear regressor.

3.3.3 NIMGCN [126].

NIMGCN is an end-to-end learning framework. The model can divided into two separate components but get trained jointly. Each component is responsible for one task, either to learn the miRNA representations or to learn the disease representations. For miRNAs, NIMGCN first constructs a homogeneous network from the pre-calculated miRNA functional similarity (**MF**). Then the model utilizes a 2 GCN layers/encoders followed by multiple nonlinear transformations or neural projections to acquire the miRNA hidden representations. The disease representation learning component employs the same architecture as that of the miRNA. However, the input network is constructed from the disease semantic similarity (**DS**).

NIMGCN treats the miRNA-disease association prediction as a matrix completion problem in which the model learns to fill in the missing entries of the association matrix **A**. The association probability corresponding to an input miRNA-disease pair (m, d) is calculated as the inner product of m's and d's latent representations. NIMGCN is trained by a customized mean squared

loss function which takes into account only the loss corresponding to the samples in the training data. NIMGCN architecture is depicted in Figure 3.2.

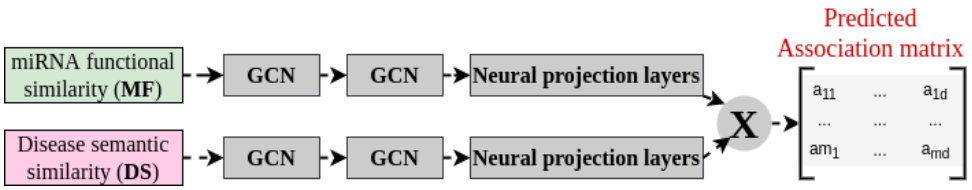


FIGURE 3.2: NIMGCN model architecture [51].

Variants of NIMGCN. We construct experiments on three of its variations in which we remove some part of the original model to construct a simpler one. The variants are summarized in Table 3.1. For NIMGCN1, we remove the GCN encoders. miRNA functional and disease similarity scores are fed directly as input to the neural projections to learn hidden representation. For NIMGCN2, we remove the stack of three neural projections. Output from GCN encoders is used directly as miRNA and disease hidden representation. And for NIMGCN3, we use only one GCN encoder and one neural projection to learn the latent features of miRNA/disease.

TABLE 3.1: NIMGCN and its variants

Model	Architecture
NIMGCN	original
NIMGCN1	without GCNs
NIMGCN2	2 GCN layers, without neural projection
NIMGCN3	1 GCN layer and 1 neural projection layer

3.3.4 DBMDA [244].

DBMDA is a two stages model which, in addition to the miRNA-disease association data, takes the pre-calculated miRNA functional similarity (**MF**, the disease semantic similarity (**DS**), and the miRNA sequence geometric similarity (**MS**) as input. For a particular miRNA m_i, its functional similarity features are the ith row in **MF**. Other types of similarity features are retrieved in a similar manner.

For a given miRNA-disease pair (m, d), DBMDA first concatenates the corresponding miRNA functional and disease semantic similarity features to form the input for an auto-encoder whose task is to extract a low dimensional representation for the input pair. At the same time, the sequence similarity features of m are fed as input to another auto-encoder to extract its latent representation. DBMDA then concatenates the output from the two auto-encoders to form the input to a rotation forest classifier whose task is to predict potential miRNA-disease associations. DBMDA architecture is illustrated in Figure 3.3.

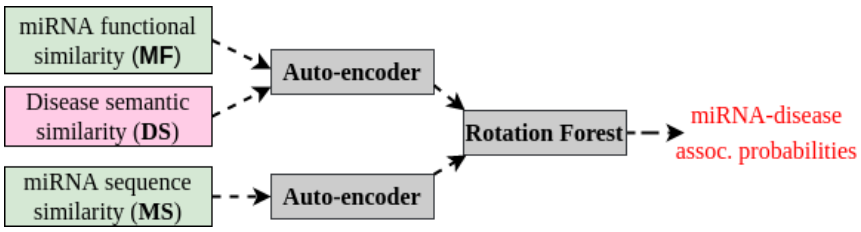


FIGURE 3.3: DBMDA model architecture [51]

Variants of DBMDA. We want to investigate the effectiveness of miRNA sequence similarity and autoencoders on DBMDA performance. Therefore, we also report results from its three variants in which we either remove the auto-encoder, the miRNA sequence similarity features, or both. As shown in Table 3.2, the DBMDA model corresponds to the original model. We obtain DBMDA1 by removing the two autoencoders which are used to extract low dimensional representations from miRNA-disease pair and miRNA sequence similarity. Instead, for a particular miRNA-disease pair, we directly concatenate the miRNA functional, disease, and miRNA sequence similarity to form a features vector. DBMDA2 corresponds to not using or removing the miRNA sequence similarity features from the input. For DBMDA3, we do not use miRNA sequence similarity features and any of the autoencoders. The concatenated features of miRNA-disease pair are directly used as input to the rotation forest classifier.

TABLE 3.2: DBMDA and its variants

Model	Autoencoder 1	Autoencoder 2	miRNA sequence features
DBMDA	✓	✓	✓
DBMDA1	x	x	✓
DBMDA2	✓	✓	x
DBMDA3	x	x	x

3.3.5 DIMIG 2.0 [162]

DIMIG 2.0 first constructs a heterogeneous network $\mathcal{G}$ in which nodes are miRNAs and PCGs and edges are derived from the known miRNA-PCG associations and PCG-PCG interactions retrieved from public databases. The model then treats the miRNA-disease association prediction as a multiclass classification problem where diseases are the labels. The miRNA nodes are labeled according to the miRNA-disease association data, while the PCG nodes are labeled according to the disease-PCG associations retrieved from a public database. DIMIG 2.0 is a semi-supervised approach that does not touch the miRNA-disease associations during training. Instead, it assumes that miRNAs are unlabeled nodes and utilizes only the PCG nodes along with their labels to learn the model parameters. As PCGs nodes are connected to miRNA nodes (according to the miRNA-PCG association data), the

learned signals are then propagated through the heterogeneous network to infer the labels for miRNAs.

3.3.6 MMGCN [205]

MMGCN is an end-to-end model that can utilize the information corresponding to four different similarity measures (the miRNA target similarity (**MP**, miRNA sequence alignment similarity (**MA**), disease semantic similarity (**DS**), and disease target similarity (**DP**). Those similarities are utilized to construct weighted input networks for representation learning. For each entity (either miRNA or disease), MMGCN employs a complex architecture consisting of two GCNs, a multichannel attention mechanism followed by a CNN layer (as illustrated in Figure 3.4) to learn its hidden representation. Similar to NIMGCN, MMGCN treats the miRNA-disease prediction as a matrix completion task. The predicted association probability for a particular miRNA-disease input pair is calculated as the inner product of the two final input representations for the corresponding miRNA and disease. MMGCN's architecture is illustrated in Figure 3.4.

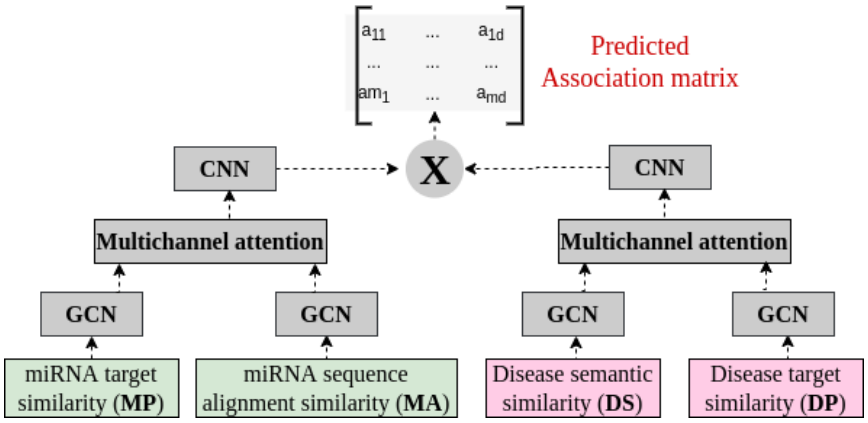


FIGURE 3.4: MMGCN model architecture.

3.3.7 GCSENET [132]

GCSENET is a two stages model which consists of a feature extractor and a binary classifier. The feature extractor serves as a preprocessing unit that utilizes a multitask learning framework to normalize the miRNA target(**MT**) and disease target features (**DT**). It takes the miRNA functional similarity (**MF**), PCG-PCG interactions, and disease semantic similarity (**DS**) as input and output predicted matrices for miRNA-PCG and disease-PCG associations. The input similarities and PCG-PCG interactions are utilized to construct networks for representation learning by GCNs. For each miRNA-PCG or disease-PCG input pair, the predicted association probability is calculated as the dot product of the two hidden representations corresponding to the two input entities.

The binary classifier takes those predicted PCG association matrices as input and employs a CNN-based classification model to accomplish the miRNA-disease association prediction task. The feature extractor and binary classifier are trained separately. Figure 3.5 illustrates GCSENET architecture.

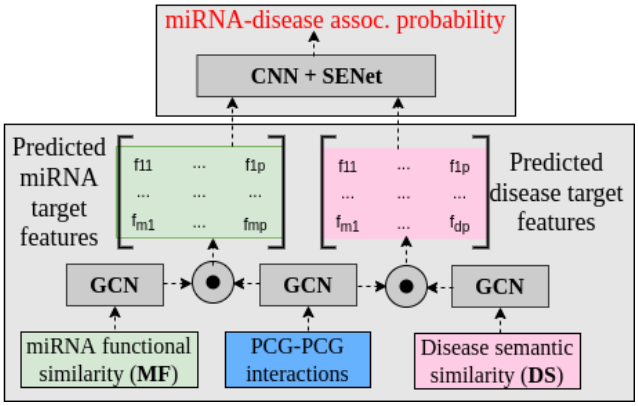


FIGURE 3.5: GCSENet model architecture.

3.3.8 NEMII [75]

NEMII is also a two stages model with two separate components: the feature extractor and binary classifier. The feature extractor is a shallow embedding learning module that takes the known miRNA-disease association to construct a miRNA-disease bipartite network for feature learning. NEMII feature extractor then utilizes a self-supervised learning model to generate a structural embedding vector for each miRNA or disease that encodes miRNA-disease association patterns extracted from the constructed bipartite network.

The classifier takes as input the learned structural embeddings (output from the feature extractor module), the miRNA family feature (**MO**), and the disease semantic similarities (**DS**). It represents each miRNA-disease input pair (m, d) as a big concatenated vector of m's and d's structural embeddings, m's family features and d's semantic similarity features (the dth row in the **DS** matrix). Then a Random Forest classifier is employed for the classification task.

Chapter 4

A consistent evaluation framework

This chapter encapsulates our first work in miRNA-disease association prediction with a detailed discussion on the limitations of existing models as well as our proposed solutions and recommendations. Most content of this chapter is taken from our conference paper: "*Towards a consistent evaluation of miRNA-disease association prediction models.*" presented at the 2020 IEEE International Conference on Bioinformatics and Biomedicine.

4.1 Limitations of existing systems and solutions

4.1.1 Data leakage problem

For a typical machine learning problem, regardless of the model architecture, the dataset used, or the underlying evaluation setup, there is/are training set(s) and its/their corresponding test set(s). A training set is used to learn the model parameters, while a test set is for evaluating the model's predictive capability on **unseen** data. That is to say, data from the test set is always hidden during the model training phase.

Data leakage in machine learning refers to a scenario in which information from the testing set is disclosed in training the model. Data leakage leads to overestimating models' predictive power and an unfair comparison between models (since different models might get affected differently). For miRNA-disease association prediction problems, the proposed models that use pre-calculated similarities that are calculated from the set of known miRNA-disease associations, like miRNA functional and GIP kernel similarities, would suffer from the data leakage problem.

To understand this, let us first consider the case of miRNA functional similarity. We note that the MISIM database is one of the most popular sources of miRNA functional similarity. The functional similarities in MISIM version 1.0 and MISIM version 2.0 are computed from the HMDD 2.0 and HMDD 3.0 databases, respectively. Those databases are later then employed as the source for most models' training and testing data.

Obviously, for a particular training-testing split, we are only allowed to use the training set information for any 'learning' operation. This implies that the miRNA functional similarities (**MF**) should be computed from the training associations solely. Such a rule is violated when the precomputed similarities are used or the similarities are not re-computed for each training

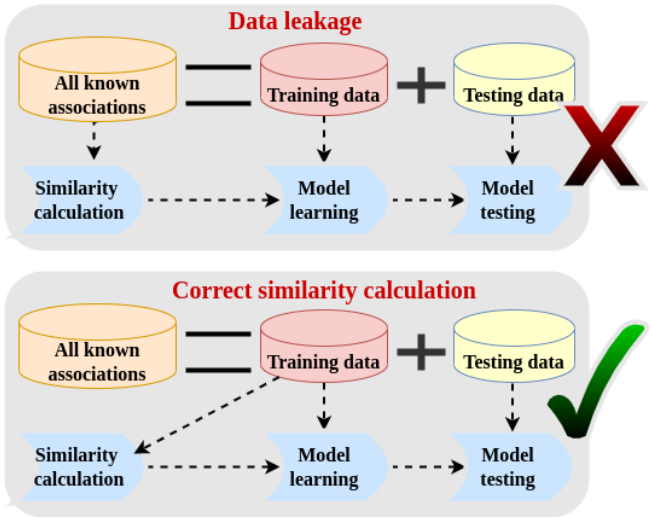


FIGURE 4.1: Data leakage problem.

set. Similarly, works using GIP similarity should also compute them using only the training associations to avoid using any information leakage from the test set. This is unfortunately not the case where most evaluation setups, though using multiple train-test splits, utilize the GIP similarities (**MG** and **DG**), which are computed just once before splitting the data. Figure 4.1 illustrates the data leakage problem.

In this work, we offer a fix to that by making available the evaluation framework (with the code) to calculate the miRNA functional, disease semantic, and GIP similarity only from the training association set. Results about different models' performance on the correctly calculated similarity measures are presented in section 4.3.

4.1.2 The evaluation setup

There are no publicly available benchmark training-testing splits for the miRNA-disease association prediction problem. In other words, there are only sets of associations between miRNAs and diseases retrieved from the HMDD databases. Each of the previous works generates its own train-test splits. K-fold-cross-validation is one of the most popular methods taken by existing approaches.

Nevertheless, even with the same data-splitting strategy, there are still many criteria that can affect the fairness of model comparison. These are (1) the use of random seeds, (2) the differences between actual training-test sets, and (3) the employed evaluation metrics.

The random seeds

Generally, many of the proposed works only run K-FoldCV once. That is to say, the author(s) run(s) only one time K-FoldCV with some random seed and then report the average performance for only that seed. It is expected that for

different random seeds, even for the same model, we will most likely get different data splits and, therefore, different results. Therefore, we believe that for K-FoldCV, the compared models should report the average performance score(s) for several random seeds for fairness purposes.

The training and testing data

In a binary classifier for the miRNA-disease association prediction problem, known miRNA-disease associations are considered positive samples. Researchers often treat unknown miRNA-disease pairs as potential negative samples. When it comes to negative samples, both the quantity and the quality of the negative sample set used in training and testing will significantly affect the model performance.

Regarding **negative training samples**, existing works either use the complete set or a random subset of potential negative samples (pairs that do not share a link) to learn the model parameters. However, using the entire set of non-connected pairs as negative training samples would result in highly imbalanced training data since the ratio of positive:negative instances are very small (around 0.029 for HMDD 2.0 dataset and around 0.039 for HMDD 3.0). A highly imbalanced training set could introduce an unwanted bias towards predicting correctly only the negative samples.

Regarding **negative testing samples**, many of the models only use a random subset of the potential negative samples set. When it comes to evaluation, from a biological perspective, people would most likely want to get a top-ranked list of miRNAs associated with a particular disease. For those systems that were evaluated only on a small subset of all possible miRNA-disease pairs, it is hard and even impossible to get such a global ranked list. Therefore, in a more realistic setting, the test set should include *all negative pairs* (excluding the ones that are already present in the train set) as against only a random subset. Such a setting is also emphasized in [235].

The evaluation metrics

The last element that should be considered is the use of evaluation metrics. In terms of measurements, most of the published articles for the current problem can use either top K Predictive Rate (topK), Area under the Receiver Operating Characteristic (AUC), or Area under the Precision-Recall Curve (AUPR).

Many existing systems use topK as the validation method for their case studies. Models were trained by the associations from one database and verified on the other. In general, the authors often take HMDD 2.0 o HMDD 3.0 as the source for training and validate the system performance on some other databases like dbDEMC 2.0 [236], miR2Disease [96] and miRCancer [231]. HMDD 3.0 and miRNA2Disease are both manually curated databases. DbDEMC 2.0 and miR2Disease are databases collected by semi-automatic methods that focus on human cancers only.

The problem lies in the fact that all those databases usually have a lot in common since they were all derived from the same set of publications related

to miRNAs and diseases. For two databases that overlap by a large portion, a highly overfitting model trained on one database could obtain a very high-performance score on the other. In that case, the topK performance score should not be used to indicate the method's predictive performance. Instead, the topK performance should be tested on a *separate test set* that does not have any intersection with the training set. The topK scores reported in section 4.3 were calculated on the corresponding testing sets only.

For AUC, as previously discussed in [235], the AUC score is sometimes misleading for highly imbalanced datasets like HMDD 2.0 and HMDD 3.0. We argue that using the Area under the Precision-Recall Curve (AUPR) is more appropriate in this case. However, a point to be noted here is the calculation of the Area under the precision-recall curve. The sklearn function to calculate the Area under the curve[1] might suffer from the overestimation of the actual area because of linear interpolation. For example, if there are just two points on x- and y-axis, it computes the Area under the line joining these two points, which is incorrect. We propose the use of Average Precision (AP) [192] which summarizes a precision-recall curve as the weighted mean of precision achieved at each threshold (the weight here is the increase in recall from the previous threshold). AP is computed as:

$$AP = \sum_{n} (R_n - R_{n-1}) \cdot P_n \tag{4.1}$$

where R_n, P_n are the recall and precision at the nth threshold correspondingly. AP does not suffer from the issues related to AUC and AUPR scores.

4.1.3 The model building problem

In this part, we are mostly concerned with models that propose either new system architectures or new types of input features. Generally, more complex systems with the addition of some non-linearity might work better but have a higher potential for overfitting. The increased set of parameters due to the models' increased complexity consequently requires more training samples to be trained/learned. For our problem of interest, the amount of positive training data is limited and is usually of very small size. Therefore, adding more complexity to the model requires a proper ablation study to ensure whether the added element results in a gain to the model performance. A similar thing should be taken into consideration when adding more features as input to the system. Larger feature spaces usually induce an increased number of parameters, which may again lead to overfitting. In this work, motivated by the idea that *simpler might be better*, besides the originally proposed models, we also construct experiments and report results on multiple simpler variants of our studied models to justify the benefits or drawbacks of the added components.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html#sklearn.metrics.auc

4.2 Data and Experimental setup

4.2.1 The benchmarked models

In this work, we study three recent state-of-the-art models which are representative of three different kinds of popular approaches:

1. NIMGCN [126] proposes a *graph convolution network* based end-to-end architecture that learns non-linear representations for miRNA and disease jointly while minimizing the prediction error.

2. DBMDA [244] uses an autoencoder-based feature extractor to learn the input representation for the binary classification model. Besides, DBMDA proposes the use of a new type of input similarity: the miRNA sequence geometric similarity (**MS**).

3. Different from the above two approaches, EPMDA [55] first extracts graph-based features from the heterogeneous miRNA-disease disease network and then uses that learned features to train a non-linear classifier.

For a thorough investigation of the models and to assess the gains obtained by using additional complexities or input features, we also experiment with several of their simpler variants. A detailed discussion of the benchmarked models and their variants has already been presented in Section 3.3.

4.2.2 Data Collection

We use two datasets derived from the HMDD 2.0 [129] and HMDD 3.0 [88] databases as our evaluation datasets. The first dataset, which from now on we denoted as HMDD2, consists of 5,430 associations between 495 miRNAs and 383 diseases. The second dataset, denoted as HMDD3, consists of 35,362 associations between 1,062 miRNAs and 893 diseases.

TABLE 4.1: The datasets statistics. n_m, n_d, N and $n_{d \notin MESH}$ denote the number of miRNAs, diseases, the known associations, and the number of diseases that are not found in MESH, respectively.

dataset	n_m	n_d	N	$n_{d \notin MESH}$
HMDD2	495	383	5,430	55
HMDD3	1,062	893	35,362	360

The MESH terms for disease semantic similarity calculation are downloaded from the National Library of Medicine[2]. Diseases are matched by names. There are 55 and 360 disease names that are not found in MESH in HMDD2 and HMDD3, respectively. For those diseases, we fill in the similarity matrix with their corresponding GIP similarity (calculated from the miRNA-disease association network) as described in Section 3.2. All miRNA

[2] http://www.nlm.nih.gov/

sequence information was retrieved from miRBase [113]. The two datasets' statistics are given in Table 4.1.

4.2.3 Experimental setup

For all the models on a particular dataset, we run five times 5-fold CV on the known miRNA-disease association set with different random seeds. Overall each model was trained/tested 25 times, corresponding to 25 possible combinations of training/testing splits. For simplicity, we call a training/testing split a data split. For all models, we use the same set of data splits. We report AUC, AP, and topK scores as mentioned in section 4.1.2. For topK, for each disease and a particular value of K, we count the number of known associations that appeared in the top K results predicted by the models (ranked by the predicted probability, the higher the probability, the higher the rank). The scores reported in Table 4.2 are the average topK results for the five popular diseases: Colon Neoplasms, Kidney Neoplasms, Prostate Neoplasms, Ovarian Neoplasms, and Lung Neoplasms.

Traing and testing data

Following previous works, we also consider all unknown miRNA-disease pairs as negative samples. Regarding the training data, we experiment with both balanced and imbalanced training sets. When it comes to testing data, we use the complete set of negative samples in the test set as described in Section 4.1.2.

Input similarities

Regarding the input similarity measures, we use the same set of similarities as proposed by the original models: GIP similarities for EPMDA (**MG** and **DG**), miRNA functional (**MF**) + disease semantic similarity (**DS**) for NIMGCN, and miRNA functional(**MF**) + disease semantic (**DS**) + miRNA sequence geometric similarity (**MS**) for DBMDA.

We resolve the data leakage issues by re-calculating all similarities according to the given data splits. More specifically, for each run, miRNA functional similarity or miRNA/disease GIP similarity scores are computed using only the training data's associations. That means when we do 5-foldCV, those similarity scores are calculated five times with different inputs. For similarities calculation, whenever possible, we use the original implementation provided by the authors. All the code for similarities computation is available at `https://git.l3s.uni-hannover.de/dong/simplifying_mirna_disease`.

Hyperparameter settings

For DBMDA, we use an autoencoder consisting of one encoder and one decoder. The encoder is a densely connected layer of size 32 (the encoded dimension explicitly given in DBMDA paper) with ReLU activation function and L1 regularization. The decoder, which has the same size as the

input, is a densely connected layer with a sigmoid activation function. We train that autoencoder for 1,000 epochs in all experiments. For the Rotation Forest [177] classifier, we use the implementation from https://github.com/digital-idiot/RotationForest with all the default parameters. For NIMGCN, we use all the author's parameters like the number of epochs, number of hidden units, etc. For EPMDA, we use the authors' code and settings for both feature calculation and the MLP regression model.

4.3 Results and discussion

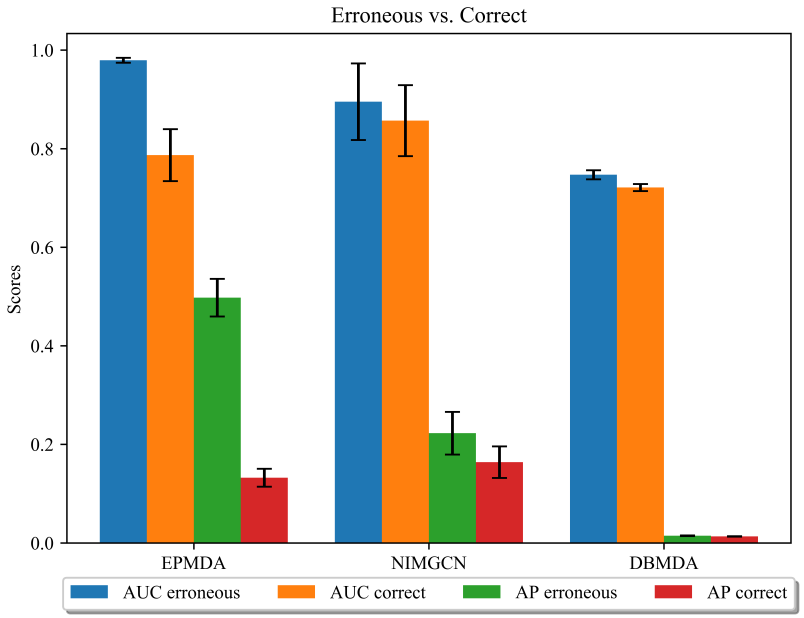


FIGURE 4.2: Erroneous and correct results on HMDD2 dataset.

In our experiments, we investigate and answer the following research questions:

- **RQ1:** Which models are most affected by the data leakage problem caused by using precomputed similarity features?
- **RQ2:** How do balanced and imbalanced training data setups compare?
- **RQ3:** How do our proposed model variants compare with the original models?

In Section 4.3.1, we answer RQ1 by comparing originally proposed models with the similarity features computed on the (i) complete association data and (ii) using only the training split. The results corresponding to RQ2 are discussed in Section 4.3.2. In Section 4.3.3, we compare the difference in the performance of different models with and without the added components (to the model architecture and the input feature set). We report the average and standard deviation of the AP/AUC scores after 25 runs corresponding to 25

data splits for all the models. However, for EPMDA and EPMDA1 on HMDD3, we could only get the results corresponding to one random seed because it took too much time (nearly 27 days for just one split) to finish the feature extraction step.

For all figures, the black lines represent the standard deviations among runs. The longer the line is, the larger the standard deviation is.

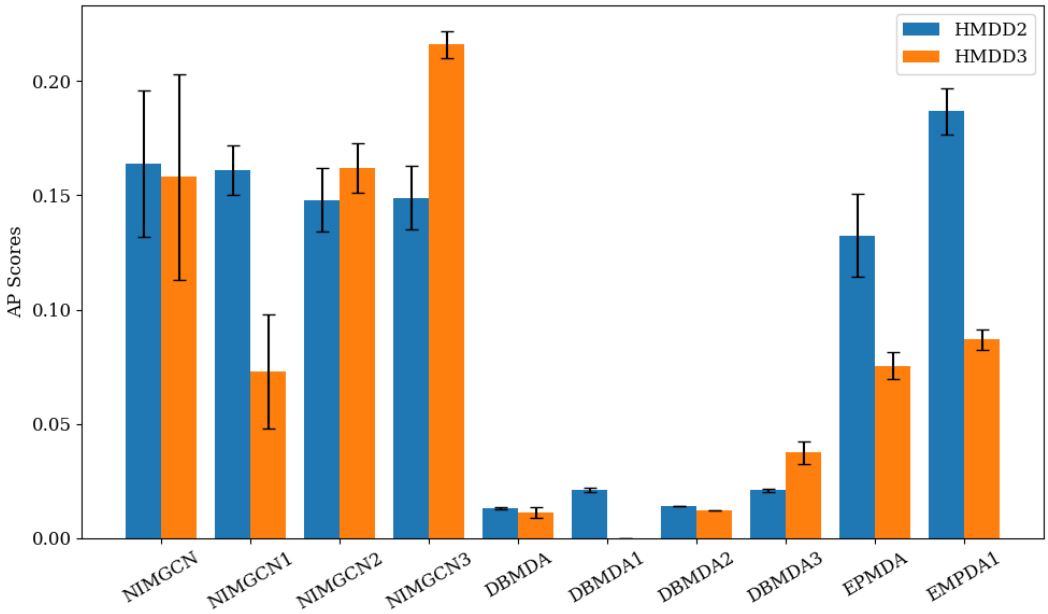


FIGURE 4.3: AP scores on HMDD2 and HMDD3 dataset with balance training set.

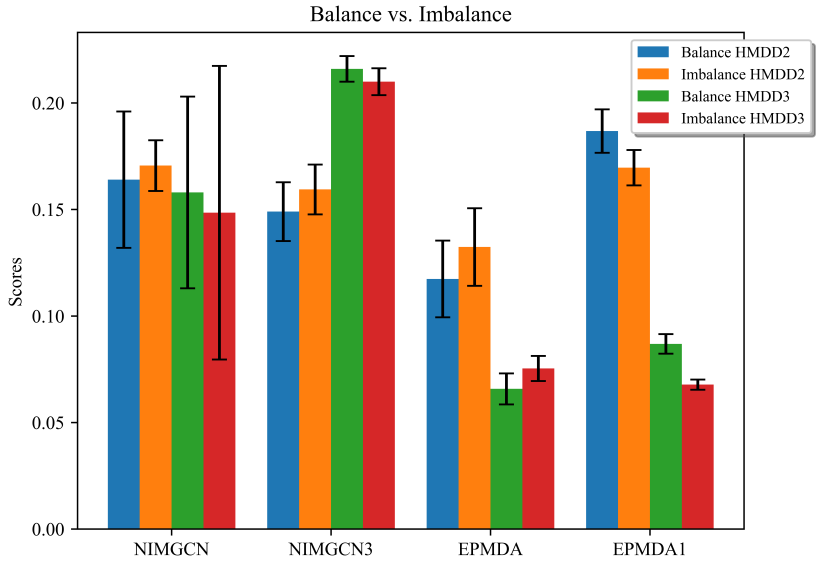


FIGURE 4.4: AP scores corresponding to balance and imbalance training data.

4.3.1 Impact of pre-computed similarities

Figure 4.2 shows the comparison of the AUC and AP scores of the three models on the HMDD2 dataset with balance training set. For all models, the "Erroneous results" denote the results retrieved from the compared models using precomputed similarities. In contrast, "Correct results" indicate the results obtained from the models that use similarities calculated at run time using only the given training data.

As expected, we find that the use of pre-calculated similarity does result in much higher AUC or AP scores. Also, different models get affected differently by the data leakage problem.

We observe the highest performance drop in EPMDA because its calculated features heavily rely on the fed GIP similarities. Note that the GIP similarities are computed using only the association information.

To represent a disease node, NIMGCN, and DBMDA use disease semantic similarity (**DS**) features, which (unlike GIP similarity) are computed from the disease ontology and not from the miRNA-disease associations. We, therefore, observe a lower impact of this correction step on DBMDA and NIMGCN.

Moreover, DBMDA also employs additional miRNA sequence similarity features that can be computed independently from the associations, making it relatively more robust to the changes in the miRNA functional similarity features. Overall, DBMDA is the least affected of all models.

4.3.2 Balanced vs. imbalanced training data

As discussed earlier, the use of an imbalance training dataset might downgrade the model performance. Figure 4.4 presents a comparison of AP scores between balance and imbalance training datasets on selected methods. The performance of DBMDA and their variants are not reported here because their performance is very low compared to other methods. Interestingly, for NIMGCN and its variants, the imbalance training set slightly boosts those models' performance in HMDD2, but it is not the case for HMDD3. In both datasets, an imbalance training dataset results in higher AP scores for EPMDA but not for EPMDA1. Since the best AP score on HMDD2 is acquired by EPMDA1 with the balance training set and the best AP score on HMDD3 is achieved by NIMGCN3 with the balance training set, we recommend the use of the balance training set.

TABLE 4.2: Average number of associations found in the TopK highest predictions for five diseases.

Method	Top10	Top20	Top30	Top40	Top50
EPMDA	1.44	3.72	6.88	10.83	15.32
EPMDA1	1.8	4.63	**8.36**	**12.83**	**17.82**
DBMDA	0.46	1.26	2.46	4.24	6.36
DBMDA1	0.53	1.29	2.53	4.06	6.01
NIMGCN	**1.98**	4.73	8.28	12.57	17.45
NIMGCN1	1.87	**4.74**	8.35	12.61	17.44

4.3.3 Impact of model architecture

To answer RQ3, we compared the three studied models and their several variants on HDMM2 and HDMM3 datasets. Figure 4.3 presents the mean AP scores for these models with standard deviation on the balance training set (except EPMDA). Table 4.2 presents the topK evaluation results corresponding to studied models and their best-performing variant on HMDD2 with the balanced training data setup. We make the following observations:

- The simpler variants show a lower standard deviation in the performance scores over multiple data splits as compared to the original models. Low variance in performance naturally increases the confidence in the model's decisions.
- Considering the average AP scores, the best performing models on both datasets is one of the proposed more lightweight variants. EPMDA1 gains the highest AP score on HMDD2 and NIMGCN3 out-performs other models by a large margin on HMDD3.
- Considering the topK evaluation, the best performing variant for HMDD2 dataset performs comparable (and sometimes better) to the original models.

From the results retrieved, we believe that adding more complexity or more input feature does not always result in a performance gain for our problem of interest. Adding more components or more information requires a proper ablation study.

4.4 Conclusion and recommendations

In this chapter, we investigate existing ML models for predicting miRNA-disease associations. We discover three issues related to many existing models that not only result in overestimating the methods' performance but also affect the fairness of model comparison and thus hinder the model development process. These include:

- The data leakage problem is rooted in the use of associations from the testing set to calculate the input features for the training phase.
- The evaluation setup is linked to the training and testing data construction, as well as the use of unreliable evaluation metrics.
- The addition of more complex architecture or input features which might lead to overfitting of models and increase in variance in model's performance

Besides presenting an in-depth study about those three types of issues, we also provide the corresponding fixes and recommendations. More particularly, we release our code to calculate the right input similarities from only the training data to overcome the data leakage problem. Additionally,

we recommend the use of a balanced training set, the complete test set with all negative pairs, and the AP score as a reliable evaluation metric. When it comes to model building, we support the construction of a careful ablation study before adding more complexity or a new type of input features to the system.

Chapter 5

The MuCoMID model

Chapter 4 analyzed the existing systems' issues as well as our proposed solutions and recommendations. Nevertheless, some open issues still exist regarding the utilization of similarity measures. *Firstly*, as these similarities are derived from the biased association data, the predictions become even more biased towards well-annotated diseases with many known associated miRNAs [87]. Moreover, the errors in the training associations will be further exaggerated in the derived feature space. *Secondly*, most similarity-based approaches cannot work effectively for new miRNAs or new diseases, i.e., instances for which no prior known associated disease (or miRNA) is available. *Thirdly*, as the code for similarity calculation is usually not publicly available, similarity-based techniques are hard to update when there are changes in the information sources employed for similarity calculation (e.g., recently discovered miRNA-disease associations, disease ontology updates, more PCG associations information, etc.).

In this chapter, we propose MuCoMiD - a multitask learning framework for the miRNA-disease association prediction problem, which can overcome most limitations in similarity-based methods. Such an approach is novel and has not been studied before. In addition, we take the lead in conducting large-scale experiments that enable a *comprehensive comparison* among models under different evaluation criteria. Besides the widely-adopted K-fold cross-validation, we propose new testing scenarios and generate new datasets to justify benchmarked models' performance on new miRNAs, new diseases, and when the training data contains many false positives. This chapter is based on our journal paper: "*MuCoMiD: A multitask graph Convolutional Learning Framework for miRNA-Disease Association Prediction*" published in IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022.

5.1 Proposed approach

In this section, we present our **Mu**ltitask graph **Co**nvolutational neural network model for **mi**RNA-**D**isease association prediction, which we refer to as MuCoMiD for brevity. Besides the miRNA-disease association prediction, MuCoMiD simultaneously learns to predict two additional side tasks. The model employs different ways of integrating domain knowledge at different stages of the learning process. In particular, information from three

biological networks: the miRNA family, PCG-PCG interaction, and disease ontology, are directly used to learn the node representations, thus, avoiding the use of pre-calculated similarities. Besides, the miRNA-PCG and disease-PCG associations are employed to construct the training data for the two side tasks.

Such added side tasks serve as regularizers and help us to incorporate the related domain knowledge. For example, a miRNA m regulates a set of proteins p that are responsible for some biological functions. Moreover, disruptions in the biological functions of p lead to certain disease condition d. Then m has some influence over disease d via p. The additional tasks of predicting miRNA-PCG associations and disease-PCG associations help us encode such influences by embedding m and d closer in the representational space. Besides, we employ an *adaptive loss balancing* technique to fine-tune the multitask loss gradients. This allows us to utilize the full power of multitask learning without resorting to exhaustive hyperparameter search.

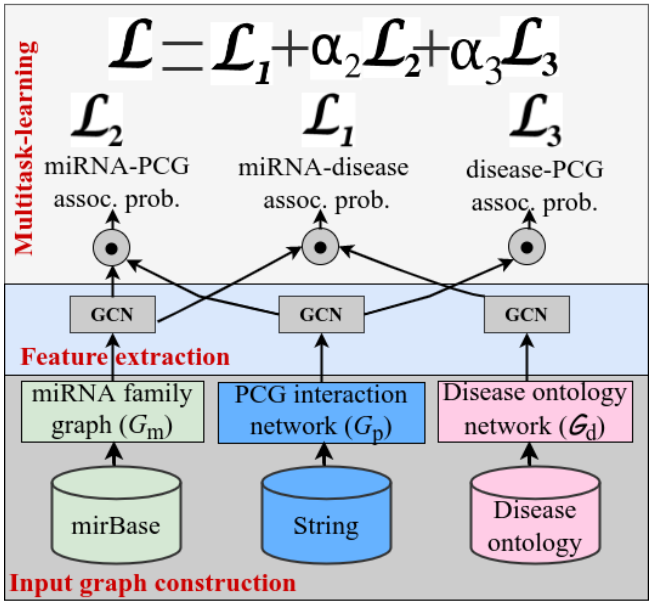


FIGURE 5.1: A schematic diagram of MUCOMID.

Figure 5.1 presents our proposed model architecture. MUCOMID consists of three main modules: (i) *input graph construction* in which we build networks corresponding to the available side information from miRNA family, PCG-PCG interactions, and disease ontology (ii) the *feature extraction* module that takes the constructed networks as input and generates the nodes' representation according to their local neighbors (iii) finally, the *multitask optimization/learning* module with one classifier for miRNA-disease association prediction, two regressors for miRNA-PCG and disease-PCG association confidence score prediction. The second and third modules get trained jointly using a multitask loss. The multitask loss is a weighted sum of the three individual task losses and is optimized using a *dynamic loss balancing* technique. In the following, we describe each module in detail.

5.1.1 Input graph construction

We start by describing the construction or retrieval of various biological networks that we leverage as additional sources of information and the corresponding rationale.

miRNA family, $\mathcal{G}_m$. A miRNA family is the group of miRNAs that share a common ancestor in the phylogenetic tree. MiRNAs that belong to the same family usually have highly similar sequence secondary structures and tend to execute similar biological functions [102]. Similar miRNAs would tend to participate in the mechanisms of similar diseases. We retrieve the miRNA family information from the mirBase database [114]. The miRNA network $\mathcal{G}_m$ is an unweighted undirected graph in which there is a connection between node A and node B if A and B belong to the same family. Figure 5.2 presents an illustration of the miRNA family network generated from our data.

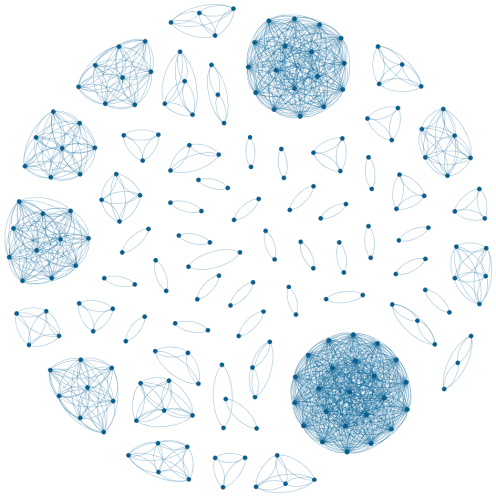


FIGURE 5.2: The miRNA family network in which each family forms a cluster.

Disease ontology, $\mathcal{G}_d$. The disease ontology [190] represents the disease etiology classes. A directed connection between two diseases exists if there exists a **is-a** relationship between them. Similar diseases can be expected to associate with similar miRNAs. The disease ontology network $\mathcal{G}_d$ is an unweighted directed network in which there is a directed connection from A to B if B is a parent of A. $\mathcal{G}_d$ can be visualized as a directed tree which contains only directed connection between children and parents nodes. Each tree layer represents one layer of abstraction. The uppermost layer represents the most general disease category. An illustration of the disease ontology is given in Figure 2.8.

PCG-PCG interaction, $\mathcal{G}_p$. PCGs interact with PCGs to carry out biological functions. Therefore, given the fact that protein-coding gene p_1 activates the expression of protein-coding gene p_2, if the miRNA m can regulate p_1 then there should be some relation between m and p_2. In other words, information from the protein-protein interaction network will bring additional

insights into the indirect relationship between miRNAs/diseases and the rest of the PCGs with which a direct interaction is not known. We download the PCG interaction data from the STRING v10.5 database [203]. As a preprocessing step, we retain only the PCG nodes that have at least one known association with miRNAs or diseases. We then divide the PCG-PCG interaction confidence scores by 1,000 to convert them to the [0,1] range and further filter out any PCG-PCG interaction with a confidence score smaller than θ_p. The results reported in section 5.3 correspond to $\theta_p = 0.3$ as it leads to the highest AP score on the NOVEL-DISEASE test set. The PCG network $\mathcal{G}_p$ is an undirected weighted network in which the edge weights are the normalized PCG-PCG interactions' confidence scores. An example of the PCG-PCG interaction network is presented in Figure 5.3 where the edge color intensity represents the interaction confidence score. The higher the score, the darker the color is.

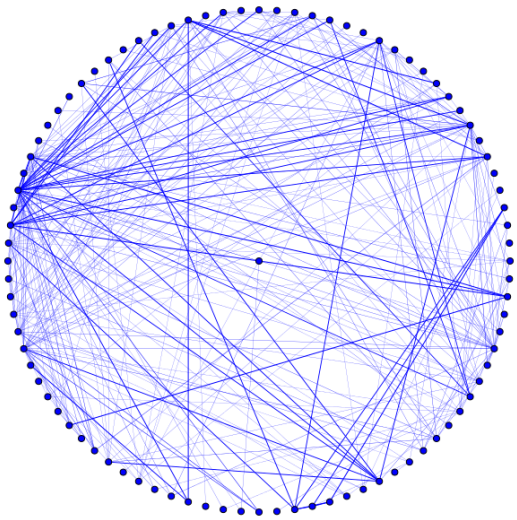


FIGURE 5.3: The PCG-PCG interaction network.

5.1.2 Feature extraction

Having constructed the relevant networks, we next extract informative node representations using the node neighborhood information. As we have no input node features, we use three embedding layers to encode the feature representation for miRNA, disease, and PCG nodes. Those embedding layers are initialized randomly and will get updated during the model training process.

An embedding layer is essentially a look-up table where the ith row corresponds to the learned representation of the ith node. The node embedding is then passed as an input feature to the graph convolutional layer. A graph convolutional layer is essentially a linear layer that transforms the node feature as an aggregation of representations of its 1-hop neighbors. In particular,

for the input adjacency matrix $\mathbf{A}$ and the node embedding matrix $\mathbf{X}$, we obtain the transformed the node feature matrix $\mathbf{X}'$ as follows:

$$\mathbf{X}' = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-1/2}\mathbf{X}\mathbf{W}, \tag{5.1}$$

where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\mathbf{I}$ is the identity matrix, $\hat{\mathbf{D}}$ is the degree matrix of $\hat{\mathbf{A}}$, and $\mathbf{W}$ is the trainable weight matrix of the graph convolutional layer. We pass the transformed representation through a ReLU activation to obtain the final representation $\mathbf{X}''$ as follows:

$$\mathbf{X}'' = max(0, \mathbf{X}') \tag{5.2}$$

As the graph semantics are different for each network, no parameter sharing is employed at this stage. We use three separate graph convolutional layers to extract the representations for miRNA, PCG, and disease nodes as illustrated in Figure 5.1. These learned representations will be fed as input to the multitask optimization/learning module explained in the next section.

5.1.3 Multitask optimization/learning

To effectively utilize information from the miRNA-PCG and disease-PCG associations, we design a multitask objective to train our model. In particular, for all input miRNA-disease, miRNA-PCG, and disease-PCG pairs, we model the pairwise representations as the elementwise products of the corresponding node features. For example, for an miRNA-disease input pair (m, d) denoted by nodes m and d, we obtain the corresponding feature vector representation as:

$$\mathbf{x}_{md} = \mathbf{X}''_m \odot \mathbf{X}''_d$$

where $\mathbf{X}''_m$ and $\mathbf{X}''_d$ correspond to the output representations of the graph convolution-based feature extraction for nodes m and d, respectively.

Using the pairwise representations, we then predict the existence of associations between miRNA-disease pairs and the confidence scores of associations for miRNA-PCG and disease-PCG pairs. In summary, we train our model with a multitask loss function calculated from these three supervised tasks and use an adaptive loss balancing technique to dynamically combine the three individual loss components at training time. Details about individual task loss and our optimization strategy are presented in the following sections.

MiRNA-disease binary classification task loss ($\mathcal{L}_1$).

We compute the probability of observing an association between an miRNA-disease input pair (m, d) as:

$$y_{md} = \sigma\left(\mathbf{w}_{MD}^T \mathbf{x}_{md}\right) \tag{5.3}$$

where $\mathbf{w}_{MD}$ is a learnable weight matrix and $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function. We use binary cross entropy to calculate the training loss for the miRNA-disease classification module as follows:

$$\mathcal{L}_1 = \sum_{m,d} -z_{md} \log y_{md} - (1 - z_{md}) \log(1 - y_{md}) \tag{5.4}$$

where z_{md} denote the target label known for the corresponding training pair.

MiRNA-PCG regression task loss ($\mathcal{L}_2$).

For an input miRNA-PCG pair (m, p), we compute the association confidence score as:

$$y_{mp} = \sigma\left(\mathbf{w}_{MP}^T \mathbf{x}_{mp}\right) \tag{5.5}$$

where $\mathbf{w}_{MP}$ is a learnable weight matrix and $\sigma(x)$ is the sigmoid function. We use the sum of squared error to calculate the training loss for the miRNA-PCG regression module as follows:

$$\mathcal{L}_2 = \sum_{m,p} (y_{mp} - z_{mp})^2, \tag{5.6}$$

where z_{mp} denotes the target confidence score.

Disease-PCG regression task loss ($\mathcal{L}_3$).

We adapt the formula presented in equation 5.5 to compute the association confidence score y_{dp} for a disease-PCG input pair (d, p). $\mathcal{L}_3$ is then calculated using the sum of squared error as in 5.6:

$$\mathcal{L}_3 = \sum_{d,p} (y_{dp} - z_{dp})^2, \tag{5.7}$$

where z_{dp} denotes the target confidence score.

Multitask optimization

We define the final loss for our model as the linear combination of three losses [123] as follows:

$$\mathcal{L} = \mathcal{L}_1 + \alpha_2 \mathcal{L}_2 + \alpha_3 \mathcal{L}_3 \tag{5.8}$$

where α_2 and α_3 are the loss weights for the two side tasks. Generally, multitask networks are difficult to train. Finding the optimal combination of individual task losses is challenging and problem-specific. A task that is too dominant during training will overwhelm the update signals and prevent the network parameters from converging to robust shared features that are useful across all tasks.

We follow the strategy presented in [123] and update α_2 and α_3 so that the difference between the two side tasks' contribution at each time step t is minimized. More specifically, at each time step t, the values for α_2, and α_3 are computed dynamically as follows:

$$\alpha_2(t) = \frac{\mathcal{L}_3(t-1)}{\mathcal{L}_1(t-1) + \mathcal{L}_2(t-1) + \mathcal{L}_3(t-1) + 10^{-10}} \tag{5.9}$$

$$\alpha_3(t) = \frac{\mathcal{L}_2(t-1)}{\mathcal{L}_1(t-1) + \mathcal{L}_2(t-1) + \mathcal{L}_3(t-1) + 10^{-10}} \tag{5.10}$$

We use an Adam optimizer [109] with a learning rate of 10^{-3} to train the multitask model.

5.2 Experimental setup

5.2.1 MiRNA-disease association data sets

We retrieve the set of miRNA-disease associations from the HMDD v2.0 database [129] and the HMDD v3.0 database [88]. As pre-processing steps, we *retain only the associations for the miRNAs and the diseases for which the PCG association information is available*. The filtered data for the HMDD v2.0 database, which from now on is denoted as HMDD2, contains 2,303 known associations between 368 miRNAs and 124 diseases. The filtered data for the HMDD v3.0 database, which from now on is referred to as HMDD3, includes 8,747 known associations between 710 miRNAs and 311 diseases. Statistics about the data are presented in Table 5.3. Note that the two datasets acquired here differ from those in chapter 4. The filtered data in this chapter do not include the miRNAs and diseases that do not associate with any PCG.

5.2.2 MiRNA-PCG association.

We obtain the miRNA-PCG associations from the RAIN database [100]. We include only the associations with the PCGs that are associated with at least one Reactome pathway [62] as these would be biologically more significant. We then normalize the association confidence scores retrieved from the database and filter out any miRNA-PCG association with a confidence score smaller than a cut-off threshold θ_m. The results presented in section 5.3 correspond to $\theta_m = 0.5$ as it results in the highest AP score for the NOVEL-DISEASE testing set. In the end, the normalized confidence scores of the retained miRNA-PCG associations are used as the target values for the miRNA-PCG association confidence score prediction side task. Statistics about the data are presented in Table 5.1.

5.2.3 Disease-PCG association.

We obtain the disease-PCG associations from the DISEASES database [169]. Here also, we retain only the associations (i) with the PCGs that are associated with at least one Reactome pathway (ii) and have the normalized confidence scores greater than or equal to a confidence cut-off threshold θ_d. The results presented in section 5.3 correspond to $\theta_d = 0.3$ as it results in the highest AP score in the NOVEL-DISEASE testing set. In the end, the normalized confidence scores of the retained disease-PCG associations are used as the target values for the disease-PCG association confidence score prediction side task.

Table 5.1 provides statistics of the three biological networks as described in section 5.1.1 and the two additional data sets described in section 5.2.2 and 5.2.3. Details about the number of miRNA-PCG, disease-PCG associations, and PCG-PCG interactions with different confidence cut-off thresholds can be found in Table 5.2, $\theta = 0$ indicates that we use the whole set without any filtering.

TABLE 5.1: Statistics for data sets with side information. $|E|$ is the number of connections/associations. n_m, n_d, and n_p are the number of miRNAs, diseases, and PCGs, respectively.

NETWORK	$\|E\|$	n_m	n_d	n_p
MIRNA-PCG	2,878	714	-	9,236
DISEASE-PCG	29,713	-	312	9,236
MIRNA FAMILY ($\mathcal{G}_m$)	1,354	217	-	-
DISEASE ONTOLOGY ($\mathcal{G}_d$)	90	-	128	-
PCG-PCG ($\mathcal{G}_p$)	1,407,590	-	-	9,236

TABLE 5.2: The number of miRNA-PCG and disease-PCG associations and PCG-PCG interactions with different confidence score cut-off threshold (θ).

	$\theta = 0$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.3$	$\theta = 0.4$	$\theta = 0.5$	$\theta = 0.6$
miRNA-PCG	178,716	69,736	23,999	5,343	4,321	**2,878**	2,112
disease-PCG	144,846	144,846	67,593	**29,713**	14,366	8,159	5,721
PCG-PCG	4,446,616	4,446,,616	2,621,256	**1,407,590**	930,930	706,144	57,6470

5.2.4 Our new testing sets

For small-size data sets like HMDD2 and HMDD3, 5-fold CV evaluation is limited as the size of the training and testing sets become much smaller. While one can use HMDD2 for training and HMDD3 for testing, such evaluation is limited as there are many overlapping associations in these two data sets. We, therefore, carefully construct the following four independent tests using the HMDD3 data set. HMDD2 *is used as the training set for evaluation with the new testing sets*. Let **M2** and **D2** be the set of all miRNAs and diseases in HMDD2, respectively. The construction of the four independent testing sets is described below.

TABLE 5.3: The miRNA-disease association data statistics where $|E|$, n_m, n_d refer to the number of associations/links, miRNAs and diseases respectively.

DATA SET	$\|E\|$	n_m	n_d
HMDD2	2,303	368	124
HMDD3	8,747	710	311
HELD-OUT1	2,669	324	110
HELD-OUT2	6,641	692	303
NOVEL-MIRNA	3,575	577	115
NOVEL-DISEASE	5,308	346	295

HELD-OUT1 for transductive testing. The HELD-OUT1 testing set contains only the associations that are present in HMDD3 but not in HMDD2. We further remove any associations involving any miRNA that is not in **M2** and any disease that is not in **D2**. By doing that, we ensure that all nodes in the testing set are partly observed during training. Finally, HELD-OUT1 contains 2,669 known associations between 324 miRNAs and 110 diseases. We randomly generate the same number of negative samples from the set of unknown miRNA-disease pairs.

HELD-OUT2 for inductive testing. We construct the HELD-OUT2 testing set by including all miRNA and disease nodes and their known associations that are present in HMDD3 but not in HMDD2. Note that different from HELD-OUT1, HELD-OUT2 might also contain the associations corresponding to the miRNA and disease nodes that are not present in the training set HMDD2. HELD-OUT2 consists of 6,641 known associations between 692 miRNAs and 303 diseases. We randomly generate the same number of negative samples from the set of unknown miRNA-disease pairs.

NOVEL-MIRNA. From the set of known associations in the HELD-OUT2 testing set, we remove any associations with the diseases that are not in **D2** to construct the NOVEL-MIRNA testing set. NOVEL-MIRNA testing set consists of 3,575 associations between 577 miRNAs and 115 diseases. Regarding the node set, NOVEL-MIRNA contains data for 253 new miRNAs that are not observed in the training set HMDD2. We randomly generate the same number of negative samples from the set of unknown miRNA-disease pairs.

NOVEL-DISEASE. Similarly, for constructing the NOVEL-DISEASE testing set, we remove any associations with the miRNAs that are not in **M2** from the HELD-OUT2 testing set. NOVEL-DISEASE contains 5,308 associations between 346 miRNAs and 295 diseases. Regarding the node set, NOVEL-DISEASE contains data for 185 new diseases that are not observed in the training set HMDD2. We randomly generate the same number of negative samples from the set of unknown miRNA-disease pairs.

A schematic Venn diagram of the four large independent testing sets is presented in Figure 5.4. The corresponding statistics are presented in Table 5.3. There are some inconsistencies between the number of miRNAs in HELD-OUT1, NOVEL-DISEASE, and **M2** or between the number of diseases in HELD-OUT2, NOVEL-MIRNA, and **D2** because for some miRNAs and diseases, all their known associations are already presented in the training set

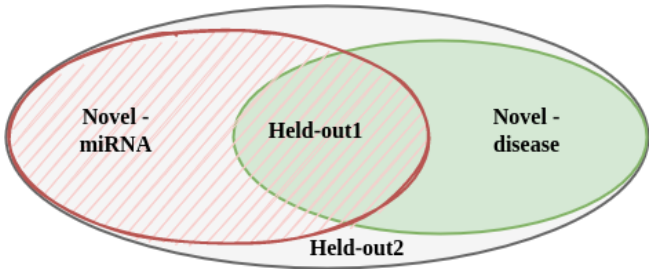


FIGURE 5.4: An illustration of the four large independent testing sets relations where HELD-OUT1 = NOVEL-MIRNA ∩ NOVEL-DISEASE, HELD-OUT2 = NOVEL-MIRNA ∪ NOVEL-DISEASE = HMDD3 \ HMDD2.

(HMDD2). Therefore, they do not appear in the known association set of the corresponding testing sets.

5.2.5 Benchmarked models

We compare our model with seven recently proposed methods: EPMDA [55], NEMII [75], NIMGCN [126], DBMDA [244], DIMIG 2.0 [244], MMGCN [205], and GCSENET [132]. More details about our benchmarked models are given in Section 3.3. Among the state-of-the-art methods, EPMDA relies on the network topology for feature extraction. NEMII learns structural embedding from the miRNA-disease network. The model also exploits the miRNA family and disease ontology information to enrich its input features. NIMGCN, DIMIG 2.0, MMGCN, and GCSENET utilize GCNs for feature learning. While DBMDA employs autoencoders for feature transformation. DIMIG 2.0, MMGCN, and GCSENET integrate similar information sources as those used by MUCOMID. Nevertheless, DIMIG 2.0 utilizes the disease-PCG associations to construct the model training objective and the miRNA-PCG associations to build the input network for feature learning. MMGCN exploits miRNA-PCG and disease-PCG associations to calculate the input similarity matrices. GCSENET uses miRNA-PCG and disease-PCG associations to construct the learning objective for its feature extractor.

5.2.6 Testing setup and evaluation

As in previous works, we perform 5-fold CV for testing on the HMDD2 and HMDD3 data sets. We run 5-fold CV with 5 random initializations. In other words, for each data set, we run each model $5 \times 5 = 25$ times and report the average performance with the standard deviation.

To test on our new testing sets, we train all models on the HMDD2 data set. We run the experiments 5 times with random initializations and report the average performance scores along with the standard deviation.

Evaluation metrics. We report the Area under the Receiver Operating Characteristic (AUC) and the Average Precision (AP) as our evaluation criteria. For our case studies, we report the number of "true" positives found at the top K highest prediction.

5.2.7 Hyperparameter settings

MuCoMiD

In all experiments, we fix the number of training epochs to 200, the embedding size and the hidden dimension both to 32. We employ Adam optimizer with a learning rate of 10^{-3} for training.

Benchmarked models

For EPMDA, DBMDA and NIMGCN, we use the code and setup released in [51]. For NEMII, MMGCN, and GCSENET, we use the same code and setup as published by the authors. We emphasize that we substantially strengthened these methods in our work since, in the original models, the authors simply take pre-calculated miRNA functional similarities from public databases (MISIM). Besides, the disease semantic similarity is originally proposed for the MESH ontology. Nevertheless, such information is not available for our disease set, and also using those pre-calculated miRNA functional similarities would lead to data leakage [51]. We instead calculate the disease semantic similarity from the disease ontology [190] and the miRNA functional similarity from the training data associations. This further points to the limited applicability of existing methods when the original information sources are updated.

For DIMIG 2.0, we use the code and parameters shared by the author. To test the model performance on our data, we compare DIMIG 2.0 with the input features as tissue expression profiles and DIMIG 2.0 with the one-hot vectors on the subset of our testing data sets that have miRNA expression profiles available. The two models acquire similar performance. This implies that the use of tissue expression profiles as node features does not affect the model performance. We can, therefore, test the model on our data for which tissue expression information is unavailable by using one-hot encoding for input node features. The results reported in Section 5.3 correspond to the model with one-hot vectors as input features on our testing data sets without removing miRNAs that do not have expression profiles.

5.3 Results

5.3.1 Results on small testing sets

Following previous works, we perform 5-fold CV experiments on the HMDD2 and HMDD3 data sets. The results are shown in Table 5.4. The testing set size for this scenario is considerably small and contains only 1/5th of the total associations. Such a train-test scenario allows us to quantify how well the models learn but is limited in testing the generalization power of the models.

MuCoMiD gains a comparable performance compared with state-of-the-art approaches. There are only minor differences in MuCoMiD's and the benchmarked methods' performance on the two data sets. Nevertheless, among the compared models, NEMII performs the best on the HMDD2 data

TABLE 5.4: Results corresponding to the 5-fold CV and transductive testing setup.

Method	HMDD2		HMDD3		HELD-OUT1	
	AUC	AP	AUC	AP	AUC	AP
EPMDA ([55])	0.744	0.783	0.520	0.594	0.427	0.49
NEMII ([75])	0.837	**0.844**	0.898	0.893	**0.705**	0.642
NIMGCN ([126])	0.785	0.803	0.795	0.800	0.623	0.601
DBMDA ([244])	0.553	0.537	0.749	0.696	0.578	0.548
DIMIG 2.0 ([162])	0.493	0.485	0.516	0.508	0.429	0.471
MMGCN [205]	0.783	0.780	0.911	**0.913**	0.682	0.627
GCSENET [132]	0.593	0.568	0.613	0.575	0.552	0.538
MUCOMID (OURS)	0.839	**0.837**	**0.916**	0.912	0.684	**0.68**
Improvement over SOTA	0.2%	-0.8%	0.5%	-0.1%	-2.9%	5.9%

set while MMGCN achieves the highest scores on the larger set HMDD3. In other words, no single benchmarked method claims its superior in both data sets. For such a reason, we claim that MUCOMID is better than all benchmarked models in the 5-fold CV testing setup.

The performance of EPMDA drops considerably for the HMDD3 data set. EPMDA learns edge features in an unsupervised manner corresponding to its contribution to a cycle of a particular length. Usually, the cycle length parameter is fixed to a small value due to an exponential increase in run time with an increase in cycle length. Moreover, the task signal is not used in learning the edge features. The loss of performance of EPMDA in HMDD3 can be attributed to the limitation of finding the best cycle length hyperparameter applicable for HMDD3. This also limits the applicability of this model to a larger variety of data sets. NEMII, NIMGCN and MMGCN perform better than DBMDA due to the higher representational capacity of the employed GCNs and the exploitation of additional graph structure information.

5.3.2 Results on small train but large test sets in transductive setting

Table 5.4 shows the results corresponding to the HELD-OUT1 testing set with a positive:negative sample rate of 1:1. Table 5.6 provides additional results regarding the larger negative sample rates. Recall that HMDD2 is used as the training set. In this scenario, the testing set size is much larger than the training set size allowing us to compare the generalization capability of the models. In general, NEMII acquires the highest AUC score, followed by MUCOMID with 2.9% lower. Regarding the AP score, MUCOMID attains the highest with a gain of at least 5.9%. This result is consistent with the experimental results for different positive:negative test sample rates presented in Table 5.6. The overall drop in performance of all models in this scenario as compared to the small testing set size cases points to the hardness of this particular testing set.

Among the benchmarked models, EPMDA and DIMIG 2.0 perform the worst, suggesting that handcrafted topology-based features extraction or semi-supervised learning on a heterogeneous network without parameter optimization are not promising approaches for the current problem. NEMII and MMGCN, which exploit multiple sources of information, gain the highest performance scores. This phenomenon further emphasizes the importance of information integration.

5.3.3 Results on inductive setting testing sets

TABLE 5.5: Results corresponding to the large inductive testing sets.

Method	NOVEL-MIRNA		NOVEL-DISEASE		HELD-OUT2	
	AUC	AP	AUC	AP	AUC	AP
EPMDA	0.44	0.529	0.5	0.5	0.417	0.513
NEMII	0.68	0.652	0.709	**0.68**	0.66	0.681
DIMIG 2.0	0.452	0.480	0.421	0.467	0.417	0.465
NIMGCN	0.533	0.519	0.672	0.666	0.534	0.509
DBMDA	0.537	0.518	0.569	0.551	0.553	0.595
MMGCN	0.556	0.504	**0.711**	0.678	0.553	0.493
GCSENET	0.543	0.518	0.557	0.536	0.557	0.517
MUCOMID	**0.701**	**0.704**	0.649	0.658	**0.667**	**0.697**
Improvement over SOTA	3.1%	8.0%	-8.7%	-3.2%	1.1%	2.3%

Table 5.5 shows the results corresponding to the inductive setting testing sets with the positive:negative sample rate of 1:1. Table 5.6 provides additional results regarding the larger negative sample rates. Recall that HMDD2 is used as the training set. Note that HELD-OUT2 is more than three times larger than the training data and contains new nodes that have not been seen in HMDD2.

In general, MUCOMID works effectively also for the inductive setting and outperforms all of its competitors on two out of the three testing sets.

Input features for benchmarked models

We note that except DIMIG 2.0, the other benchmarked models have some issues with generating predictions in the inductive settings. Specifically for EPMDA, which relies on the Gaussian Interaction Profile kernel similarities extracted from the known associations, input features for new miRNA and diseases will be all zeros. NEMII, which concatenates the extracted features from the miRNA-disease association network with the miRNA family and the disease semantic similarity features, will have part of its input features for new miRNAs or new diseases as random values. Likewise, the miRNA functional similarity for new miRNAs in the MMGCN, NIMGCN, and DBMDA, GCSENET models will be all zeros. Therefore, a part of the miRNA-disease pairs' final input representation in those models will be random values.

The NOVEL-MIRNA and HELD-OUT2 testing sets

These two testing sets contain known associations for hundreds of new miRNAs, which are not observed in the training data set.

MUCOMID vs. DIMIG 2.0. Though DIMIG 2.0 can predict the association probabilities for new miRNAs and new diseases, the differences in the two models' architecture and learning objectives lead to a significant difference in their performance. Unlike MUCOMID and other state-of-the-art models, DIMIG 2.0 is a semi-supervised method that uses only disease-PCG associations during training but not the known miRNA-disease associations. Also, DIMIG 2.0 is formulated as a multi-label classification problem with large but very sparse label matrices. The high sparsity of the labels, along with the high class imbalance, leads to a degradation in learning.

MUCOMID vs. other methods. As discussed in section 5.3.3, EPMDA, NEMII, NIMGCN, MMGCN, DBMDA, and GCSENET have a part of the input features corresponding to new miRNAs to be zeros or random values. For such reasons, the performance of state-of-the-art methods drops significantly on those two testing sets. MUCOMID significantly outperforms all of its competitors with a gain of up to 8% in AP score. Though the structural embeddings for new miRNAs and new diseases in the NEMII model are random, its performance still ranks the second-highest, suggesting that the miRNA family and disease semantic similarity features are quite informative for the current classification problem.

The NOVEL-DISEASE testing set

The NOVEL-DISEASE testing set contains known associations for 185 diseases that are not observed during training. For this testing set, MUCOMID is outperformed by NEMII and MMGCN by small margins. We argue that the direct use of miRNA-disease association training data to compute miRNA similarity or structural embedding by NIMGCN, MMGCN, and NEMII leads to their better performance (on the NOVEL-DISEASE data set) than MUCOMID. These models are usually biased towards giving high scores to the well-known miRNAs (for which a lot of association information is already known in training data), leading to overall better scores.

Performance on the testing sets with more negative samples

This section presents the results corresponding to state-of-the-art methods' performance on large independent testing sets(as described in Section 4.4 in the main paper) with different positive:negative rates. We vary the positive:negative samples rate such that it is one of {1:1, 1:5, 1:10}.

Table 5.6 presents the results for MUCOMID and state-of-the-art models on the new data. For each positive sample set, for each negative test rate, we randomly sample 10 negative sample sets with different seeds. Also, for each positive + negative set combination, we run each model 10 times to mitigate the effect of parameter initialization. The results reported in Table 5.6 is the

TABLE 5.6: AP scores on large test sets with different positive:negative sample rates. nr_1, nr_5, and nr_{10} correspond to the positive:negative rate of 1:1, 1:5, and 1:10.

Data	HELD-OUT1			NOVEL-MIRNA			NOVEL-DISEASE			HELD-OUT2		
	nr_1	nr_5	nr_{10}	nr_1	nr_5	nr_{10}	nr_1	nr_5	nr_{10}	nr_1	nr_5	nr_{10}
EPMDA	0.494	0.164	0.09	0.497	0.171	0.095	0.501	0.168	0.092	0.497	0.169	0.093
NEMII	0.727	0.368	0.233	0.754	0.408	0.269	0.721	0.365	0.23	**0.741**	**0.399**	**0.262**
NIMGCN	0.753	0.397	0.255	0.517	0.18	0.099	**0.771**	**0.432**	**0.286**	0.482	0.163	0.089
DBMDA	0.629	0.266	0.155	0.58	0.223	0.127	0.625	0.262	0.153	0.579	0.225	0.129
MMGCN	0.756	0.404	0.261	0.54	0.194	0.108	0.749	0.404	0.262	0.507	0.177	0.098
MUCOMID	**0.78**	**0.457**	**0.314**	**0.771**	**0.445**	**0.303**	0.737	0.398	0.263	0.73	0.395	**0.262**

average of 1000 runs for each model (with 10 different negative training sets and 10 different negative testing sets. For each set, each model gets tested 10 times).

We do not have the results available for the GCSENET model because the originally released code is not optimized to run on GPU. With CPU, it can only run on one of our CPU-intensive servers and takes around 12 hours to finish one run. It would take us 12x100x3 = 3600 hours to finish, which is infeasible.

The reported results show similar trends as what is already presented above. MUCOMID significantly outperforms state-of-the-art models on two out of the four testing sets. On the HELD-OUT2 testing set, its performance closely follows the best method (NEMII). On the NOVEL-DISEASE testing set, MUCOMID's performance ranked the third, higher than NEMII. We note that no single method consistently outperforms MUCOMID on both the NOVEL-DISEASE and HELD-OUT2 testing sets. These results again highlight the superiority of our proposed model and the importance of integrating information from multiple sources.

5.3.4 Ablation study

Multitask vs. Single task

We conduct an ablation study to analyze the contribution of the additional tasks. The single task baseline (MUCOMID-STT) employs a similar architecture as that of MUCOMID but without the miRNA-PCG and disease-PCG association confidence score prediction side tasks. In other words, it also learns miRNA and disease representation from the miRNA family and the disease ontology networks, respectively. However, MUCOMID-STT only has one classifier layer for the miRNA-disease association prediction task, instead of one classifier and two regressors as that of MUCOMID.

Table 5.7 presents the results for MUCOMID and MUCOMID-STT on the both 5-fold CV and the independent testing setup. MUCOMID performs comparably to its single task variant on the 5-fold CV testing setup while significantly supersedes its competitor on all of the large independent testing sets with *much less standard deviation values among runs*. These results are

even more significant when considering the size of the testing data. The performance gain highlights the contribution of the two added side tasks. Since PCGs are the most important links between miRNAs and their associated diseases [153], miRNA-PCG and disease-PCG prediction tasks also bring additional insights into the miRNA-disease association prediction problem.

TABLE 5.7: Multitask vs. single task ablation study results.

Method	MUCOMID		SINGLE TASK MUCOMID	
	AUC	AP	AUC	AP
HMDD2	0.839 ± 0.012	0.837 ± 0.015	$\mathbf{0.843 \pm 0.01}$	$\mathbf{0.843 \pm 0.016}$
HMDD3	$\mathbf{0.916 \pm 0.005}$	0.912 ± 0.006	$\mathbf{0.916 \pm 0.004}$	$\mathbf{0.913 \pm 0.005}$
HELD-OUT1	$\mathbf{0.684 \pm 0.003}$	$\mathbf{0.68 \pm 0.005}$	0.674 ± 0.094	0.663 ± 0.092
Novel-miRNA	$\mathbf{0.701 \pm 0.002}$	$\mathbf{0.704 \pm 0.002}$	0.683 ± 0.095	0.689 ± 0.096
Novel-disease	$\mathbf{0.649 \pm 0.005}$	$\mathbf{0.658 \pm 0.006}$	0.606 ± 0.084	0.618 ± 0.086
HELD-OUT2	$\mathbf{0.667 \pm 0.006}$	$\mathbf{0.697 \pm 0.007}$	0.618 ± 0.086	0.66 ± 0.092

Model architecture

TABLE 5.8: Model architecture ablation study results. n_{r1}, n_{r5}, n_{r10} correspond to the positive:negative rates of 1:1, 1:5, and 1:10.

Data	MUCOMID			MUCOMID-GAT			MUCOMID-lin			MUCOMID-2GCN		
	nr_1	nr_5	nr_{10}	nr_1	nr_5	nr_{10}	nr_1	nr_5	nr_{10}	nr_1	nr_5	nr_{10}
HELD-OUT1	.78	.457	.314	**.783**	**.459**	**.315**	.778	.452	.307	.778	.448	.302
NOVEL-MIRNA	.771	.445	.303	**.777**	**.46**	**.318**	.75	.411	.272	.763	.425	.28
NOVEL-DISEASE	**.737**	**.398**	**.263**	.722	.369	.235	.727	.379	.243	.718	.355	.22
HELD-OUT2	**.73**	**.395**	**.262**	.725	.385	.253	.709	.36	.23	.706	.342	.211

This section provides a more in-depth ablation study to justify MUCOMID's choice of architecture. We compare MUCOMID with three of its variant: (1) MUCOMID without ReLU activation (MUCOMID-lin) which is a linear model without any non-linearity added; (2) MUCOMID with Graph Attention network (MUCOMID-GAT) in which we replace the GCN layer with a GAT layer; and (3) MUCOMID-2GCN: MUCOMID with 2 GCN layers and a ReLU activation added in between.

Table 5.8 presents the acquired scores for the benchmarked models on the testing sets described in Section 5.3.3. Compared with MUCOMID-lin and MUCOMID-2GCN, MUCOMID acquires comparable results on the transductive testing data set (HELD-OUT1) and significantly gains higher AP scores in the remaining independent testing sets. The gains are more significant when there are new miRNAs and/or new diseases.

Compared with MUCOMID-GAT, MUCOMID acquires comparable performance on the transductive testing set while significantly outperforming its competitor on two of the three inductive testing sets. These results claim that the GCN architecture employed by MUCOMID is more appropriate than GAT for the miRNA-disease association prediction task, given the limited available training data.

To summarize, MuCoMiD presents a good balance between model complexity and performance. It performs better than a linear GCN. Adding more complexity by using GAT does not show considerable improvements.

MuCoMiD performance regarding different side information sources

TABLE 5.9: MuCoMiD performance with different PCG-PCG data sources.

Database	Held-out1		Novel-miRNA		Novel-Disease		Held-out2	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP
STRING v10.0	0.681	0.677	**0.704**	**0.705**	0.627	0.646	0.647	0.687
STRING v10.5	**0.684**	**0.68**	0.701	0.704	**0.649**	**0.658**	**0.667**	**0.697**
STRING v11.5	0.681	0.678	0.702	**0.705**	0.617	0.633	0.637	0.674

Since the data sources employed for miRNA-PCG and disease-PCG associations only have one version so far, we cannot construct an ablation study with different versions of such data. In Table 5.9, we present the results corresponding to different versions of the STRING database. More specifically, we report the performance of our model on the testing sets described in Section 5.3.3 corresponding to STRING v10.0, v10.5 (the current version employed by our model), and v11.5. Each entry in Table 5.9 is the average after 1000 runs. We observe that MuCoMiD's performance also varies according to the input PCG-PCG interaction set. Nevertheless, the difference is not too significant in the transductive settings (Held-out1) and the Novel-miRNA testing set. We observe more significant differences for the Novel-Disease and Held-out2 testing sets. It could be the case that the PCG interaction profile contributes more toward the model decision for new diseases.

5.4 Case studies

5.4.1 The Parkinson disease case study

Parkinson disease (PD) is the second most common neurodegenerative disease worldwide [115]. Existing human association studies for Parkinson disease resulted in inconsistent findings with several "false positives" reported by [191].

In this case study, we aim to answer the question of "*How effectively can* MuCoMiD *help in identifying false positives?*". Towards that, we manually construct a "gold standard" data set based on the data deposited in the HMDD databases and the data collected from [191]. We mark 12 miRNAs as "true positives" (those that are confirmed as true positives in the meta analysis [191]), 33 miRNAs as "false positives" (which are marked as positives in the HMDD databases but are confirmed as negatives in the meta-analysis) and 116 miRNAs as "true negatives" (those that are confirmed as negative by the meta-analysis). Note that among the 12 true positive miRNAs, only 8 are marked as positive in the HMDD databases.

Training and Testing data setup. We first construct the training data set by including diseases other than Parkinson. Let $\mathbf{H} = \text{HMDD2} \cup \text{HMDD3}$. We remove any known associations for Parkinson disease from $\mathbf{H}$ to obtain the $\mathbf{H}'$ data set. As for Parkinson alone, the false positive rate is nearly 3 fold. We expect such a high number of false positives also for other diseases. To mitigate the effect of high false-positive rates for other diseases, we construct $\mathbf{H}^*$ from $\mathbf{H}'$ such that each miRNA only associates with μ diseases where μ is the average number of diseases associated with a particular miRNA in $\mathbf{H}'$. If the number of diseases associated with miRNA m_1 is larger than μ, we randomly sample μ diseases from the set of known associated diseases. If the number of associated diseases for m_1 is smaller than μ, we sample with duplicates μ diseases from the set of known diseases. We follow the same strategy to obtain the negative samples set for $\mathbf{H}^*$.

Next, we create the training subset corresponding to the Parkinson's disease associations, which we refer to as $\mathbf{P}_{train}$. The number of positive samples in $\mathbf{P}_{train}$ consists of 8 miRNA-Parkinson associations with 8 "true positive" miRNAs that appear in $\mathbf{H}$ and $FP \times 8$ false positives Parkinson-miRNA associations. In our experiment, FP is varied over: {0.25, 0.5, 0.75, 1, 2, 5, 10}. The negative samples in $\mathbf{P}_{train}$ consist of every possible combination between Parkinson and any miRNAs that are not "true positive" or "false positive".

The Parkinson's training data is the union of $\mathbf{H}^*$ and $\mathbf{P}_{train}$. The Parkinson's testing data consists of all pairwise combinations between Parkinson and 12 "true positive" and 116 "true negative" miRNAs identified in [191].

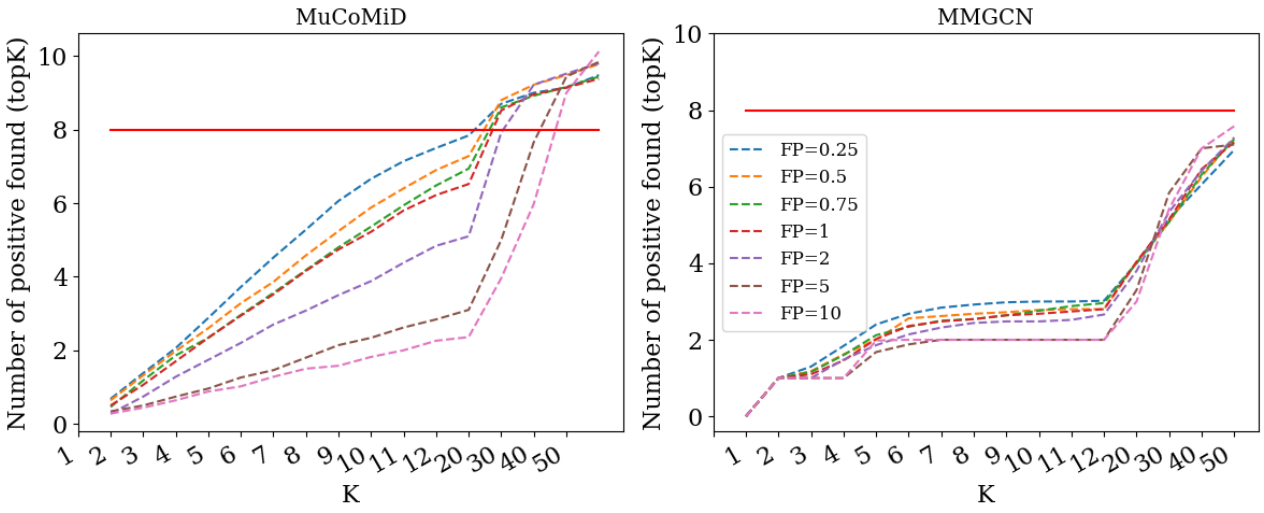


FIGURE 5.5: The Parkinson disease case study results.

Figure 5.5 presents the number of true positives found in the top K highest predictions of MUCOMID and the MMGCN model on the Parkinson testing set with varying false-positive sample rates. 'FP' refers to the false positive:true positive sample rate in the training data. The dense red line represents the number of true positives in the training data. Note that for each false positive rate, we run each model with 10 different sampled sets,

and for each set, we run the model 5 times to make the comparison as fair as possible.

With the increase in the false positive rate, there is a decrease in the performance of both the models, which is intuitive. Remarkably, MUCOMID significantly outperforms its competitor in differentiating between the "true positives" and the "false positives". "True positive" miRNAs consistently appear in the top predictions generated by MUCOMID. While for MMGCN, even with only 25% "false positives" added to the training set, for K=8, there are only 2.92 "true positives" (on average) found. On the other hand, the average number of true positives (for K=8) found by MUCOMID is 6.06. These results highlight the added benefit of our proposed model. We believe that the two added side tasks help inform the model and prevent it from overfitting the noisy training association data. At the same time, it further validates our concerns associated with secondary features-based methods. MMGCN, which extracts the input features from the training associations, cannot well differentiate the false positives from the true positives. We observe similar results in comparison with other baselines.

5.4.2 Case studies concerning well-studied diseases

To further demonstrate our multitask model's predictive capability, we evaluate the model for three specific diseases: BREASTCANCER, PANCREATICCANCER, and DIABETESMELLITUS-type 2. By constructing these case studies, we showcase our model's predictive performance in predicting associations for a specific new disease. To do that, for a specific disease d, we select the pairs associated with d from $\mathbf{H}$ = HMDD2 $\cup$ HMDD3 to use as the testing set. The remaining pairs in $\mathbf{H}$ are used as the training data. We do negative sampling for both the training and testing set so that the number of positive and negative samples in both training and testing sets are equal. For the testing set, the negative pairs are generated corresponding to only the disease d. The data set statistics for our case studies are presented in Table 5.10 where we use the disease names to denote our generated data sets.

TABLE 5.10: The case studies' data statistics, where n^+ and n^- refer to the number of positive and negative associations, respectively.

DISEASE	TRAIN SAMPLES		TEST SAMPLES	
	n^+	n^-	n^+	n^-
BREASTCANCER	8423	8423	324	324
PANCREATICCANCER	8578	8578	169	169
DIABETESMELLITUS	8640	8640	107	107

Figure 5.6 presents the topK evaluation results, while Table 5.11 shows the top 50 most confident predictions of the associated miRNAs for the three diseases. For each case study, the known association for that particular disease is completely hidden from the model training process. The statistics of our training and testing data can be found in Table 5.10.

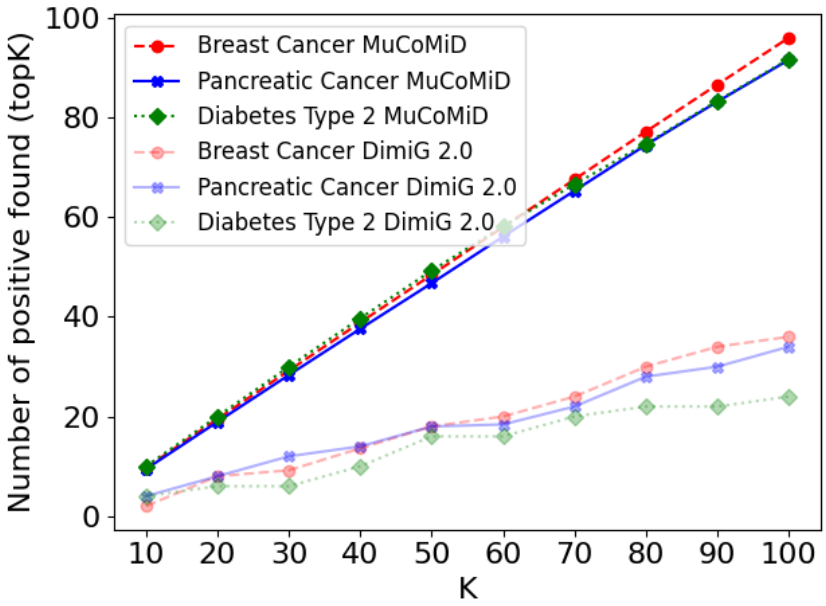


FIGURE 5.6: Results for case studies concerning well-studied diseases.

Looking at Figure 5.6, though the case studies' diseases are completely new, our model still attains near-perfect predictions for the top 40 predictions. Compared with DIMIG 2.0 for the top 50 predictions, our model has a gain of at least 160%.

For BREASTCANCER, the model acquires $\sim$ 96.04% accuracy for the top 100 predictions. For DIABETESMELLITUS-type 2, the number of known positive associations is only 107, but out of the top 100 predictions, we can correctly recognize 91.76 associations (this number is the average of the number of found associations over five runs). These results again claim the effectiveness of our model in predicting potential associations for new diseases.

5.5 Conclusion

We propose a multitask graph convolutional learning framework, MUCOMID for the problem of predicting miRNA-disease associations. Our end-to-end learning approach allows automatic feature extraction while incorporating knowledge from five heterogeneous biological information sources. Incorporating multiple sources of information helps compensate for the lack of information in any single source and, at the same time, enables the model to generate predictions for any new miRNA or disease. Unlike previous works, our model can be employed in both transductive and inductive settings. To test the generalization power of models, we test them on both the existing benchmarked setup and on our constructed large independent testing sets. Large-scale experiments in several testing scenarios highlight the superiority of our approach. An ablation study is added to highlight the side tasks' contribution. We release all the code and data used in this study for reproducibility and future research at `https://git.l3s.uni-hannover.de/dong/cmtt`.

We believe that our design principles will be of independent interest to other biomedical applications where data scarcity is a major challenge. In particular, the use of multitask learning to integrate information from heterogeneous information sources to overcome the problems of data scarcity and unreliability of one single data type is a unique perspective and has not been studied for computational problems in biomedicine.

TABLE 5.11: The top 50 miRNAs that have the highest association probabilities with the case study diseases produced by MUCOMID (average after five runs).

No.	BREASTCANCER		PANCREATICCANCER		DIABETESMELLITUS-type 2	
1.	hsa-mir-21	0.914	hsa-mir-21	0.959	hsa-mir-21	0.841
2.	hsa-mir-155	0.91	hsa-mir-146a	0.953	hsa-mir-146a	0.821
3.	hsa-mir-146a	0.909	hsa-mir-126	0.95	hsa-mir-155	0.818
4.	hsa-mir-145	0.907	hsa-mir-34a	0.95	hsa-mir-145	0.814
5.	hsa-mir-126	0.907	hsa-mir-146b	0.95	hsa-mir-29b	0.809
6.	hsa-mir-146b	0.906	hsa-mir-34b	0.95	hsa-mir-126	0.808
7.	hsa-mir-34b	0.905	hsa-mir-222	0.949	hsa-mir-34a	0.806
8.	hsa-mir-34a	0.905	hsa-mir-221	0.949	hsa-mir-221	0.804
9.	hsa-mir-222	0.902	hsa-mir-155	0.949	hsa-mir-34c	0.8
10.	hsa-mir-34c	0.902	hsa-mir-34c	0.948	hsa-mir-222	0.799
11.	hsa-mir-29b	0.902	hsa-mir-29b	0.948	hsa-mir-142	0.796
12.	hsa-mir-221	0.899	hsa-mir-145	0.948	hsa-mir-150	0.79
13.	hsa-mir-31	0.898	hsa-mir-31	0.947	hsa-mir-223	0.785
14.	hsa-mir-142	0.895	hsa-mir-27b	0.945	hsa-mir-205	0.784
15.	hsa-mir-210	0.894	hsa-mir-205	0.945	hsa-mir-144	0.783
16.	hsa-mir-144	0.892	hsa-mir-181a	0.945	hsa-mir-122	0.779
17.	hsa-mir-223	0.891	hsa-mir-27a	0.943	hsa-mir-214	0.779
18.	hsa-mir-150	0.891	hsa-mir-214	0.942	hsa-mir-27a	0.773
19.	hsa-mir-122	0.887	hsa-mir-150	0.942	hsa-mir-27b	0.771
20.	hsa-mir-205	0.887	hsa-mir-142	0.942	hsa-mir-26a	0.77
21.	hsa-mir-26a	0.883	hsa-mir-210	0.941	hsa-mir-29a	0.766
22.	hsa-mir-27b	0.882	hsa-mir-223	0.941	hsa-mir-375	0.76
23.	hsa-mir-27a	0.881	hsa-mir-122	0.94	hsa-mir-24	0.758
24.	hsa-mir-29a	0.877	hsa-mir-26a	0.938	hsa-mir-192	0.755
25.	hsa-mir-214	0.875	hsa-mir-29a	0.935	hsa-mir-143	0.754
26.	hsa-mir-143	0.874	hsa-mir-375	0.934	hsa-mir-92a	0.751
27.	hsa-mir-29b-1	0.874	hsa-mir-29c	0.933	hsa-mir-9	0.749
28.	hsa-mir-29c	0.873	hsa-mir-92a	0.933	hsa-mir-19b	0.747
29.	hsa-mir-375	0.872	hsa-mir-196a	0.932	hsa-mir-486	0.745
30.	hsa-mir-24	0.872	hsa-mir-192	0.932	hsa-mir-206	0.743
31.	hsa-mir-29b-2	0.869	hsa-mir-29b-1	0.932	hsa-mir-23b	0.736
32.	hsa-mir-92a	0.867	hsa-mir-486	0.932	hsa-mir-183	0.731
33.	hsa-mir-9	0.865	hsa-mir-342	0.931	hsa-mir-23a	0.729
34.	hsa-mir-23b	0.865	hsa-mir-23b	0.928	hsa-mir-1	0.725
35.	hsa-mir-124	0.863	hsa-mir-143	0.927	hsa-mir-17	0.725
36.	hsa-mir-19b	0.863	hsa-mir-23a	0.927	hsa-mir-20a	0.724
37.	hsa-mir-206	0.863	hsa-mir-24	0.926	hsa-mir-93	0.724
38.	hsa-mir-342	0.863	hsa-mir-16	0.926	hsa-mir-18a	0.724
39.	hsa-mir-16	0.863	hsa-mir-9	0.925	hsa-mir-182	0.723
40.	hsa-mir-486	0.862	hsa-mir-125b	0.922	hsa-mir-19a	0.723
41.	hsa-mir-196a	0.861	hsa-mir-139	0.921	hsa-mir-18b	0.722
42.	hsa-mir-181a	0.858	hsa-mir-206	0.92	hsa-mir-20b	0.722
43.	hsa-mir-183	0.857	hsa-mir-215	0.919	hsa-mir-215	0.712
44.	hsa-mir-192	0.857	hsa-mir-367	0.918	hsa-mir-22	0.711
45.	hsa-mir-23a	0.856	hsa-mir-128	0.916	hsa-mir-15a	0.709
46.	hsa-mir-139	0.855	hsa-mir-182	0.916	hsa-mir-195	0.707
47.	hsa-mir-125b	0.852	hsa-mir-186	0.915	hsa-mir-125b	0.701
48.	hsa-mir-186	0.852	hsa-mir-19a	0.911	hsa-mir-15b	0.701
49.	hsa-mir-1	0.852	hsa-mir-183	0.911	hsa-mir-96	0.694
50.	hsa-mir-182	0.851	hsa-mir-20a	0.907	hsa-mir-128	0.691

Chapter 6

The MPM model

In Chapter 5, we propose a novel multitask learning framework that can address most limitations in existing works for miRNA-disease association prediction. Nevertheless, regarding the incorporated data sources, MUCOMID still applies naive filtering techniques with multiple hard thresholds to remove data redundancy. Such an approach requires a time-consuming parameter fine-tuning process each time the incorporated data sources change. Besides, since such filtering operation does not rely on any biological ground, the outcome is also questionable. In this chapter, we propose a biologically data-driven approach that can effectively address such an issue by a simple yet effective mechanism to enrich and filter the incorporated data. With the updates in data sources, data preprocessing, and model learning strategy, we are the first to be able to generate predictions for most known miRNAs and diseases. Our contribution also lies in developing and releasing a publicly available, easy-to-use website that encapsulates all the generated predictions along with their corresponding biologically relevant information to foster assessments and adoption. This chapter is based on our journal paper: "*A message passing framework with Multiple data integration for miRNA-Disease association prediction*" published in Scientific Reports, 2022.

6.1 Method

We propose MPM- a biologically-motivated data-driven approach that can overcome most of the issues persisting in existing works. In addition to fusing multiple knowledge sources, we propose a parameter-free mechanism to enrich and control the quality and quantity of the added data. A crucial design decision of our approach includes modeling the biological relevance of miRNAs for a particular disease via the associated PCGs. We model each miRNA or disease as a directed network built from the miRNA-PCG, disease-PCG associations, and PCG-PCG functional interactions. MPM employs a message passing framework operating over the constructed networks to enrich the existing data with potential missing links or indirect connections.

To overcome the noisy data problem, we employ a *feature selection* strategy with a side-supervised task generated from the well-annotated MESH ontology [15]. Feature selection at this stage allows us to reduce the tens of thousands of associated PCGs to only one hundred most important PCGs.

This enables us to control the quality and the quantity of the added PCG-related information without introducing any additional parameters. This is extremely important, especially in the context of learning from scarce data when over-parameterized models can easily overfit.

Next, we encapsulate the enriched and filtered PCG connections into the existing miRNA-disease bipartite network to overcome the isolated nodes problem in existing works. Since PCGs are important connections between miRNA and diseases [153], the patterns learned from the miRNA-PCG-disease interconnected networks should be a rich source of information for the miRNA-disease association prediction problem. At the same time, the newly introduced heterogeneous network will include biological connections between new miRNAs or new diseases and their associated PCGs. The learning signals will thus transfer from known miRNAs or known diseases to the new miRNAs or new diseases via the PCGs. We employ the SDNE model to extract the patterns (or pre-trained embeddings) from the constructed heterogeneous network. Besides the structural features, the final miRNA-disease pair representation is further augmented with information from the miRNA family and disease semantic similarity and then fed as input to a Random Forest classifier to perform the association prediction task.

A schematic diagram of MPM with its main components is presented in Figure 6.1. MPM consists of a message passing layer (section 6.1.1) , a feature selection with a side supervised task (section 6.1.2), a Structural Deep Embedding network (section 6.1.3), and a binary classifier (section 6.1.4). We use gray for the model's components/modules, green and pink for miRNA and disease-related components, respectively.

6.1.1 The message passing framework/module

The data sources

Table 6.1 provides the statistics for our employed data sources. In the following, we describe each source in detail and present the information corresponding to how we utilize it.

The protein functional interaction network. Protein coding genes (PCGs) are essential connections between miRNAs and diseases [153]. MiRNAs can affect the PCG transcriptions, resulting in protein expression changes, which can then lead to diseases. Therefore, besides the knowledge about the protein-protein interactions as already exploited by MuCoMiD [52] (ref. Section 5.1.1), the knowledge related to whether a particular protein can regulate/inhibit/-catalyze/activate another protein is also very important for the miRNA-disease association prediction task. We refer to the multi-relational protein-protein interaction network, where an edge corresponds to a protein functional relation as *protein functional interaction network*.

A pictorial example of the protein functional interaction network is presented in Figure 6.2. Different relations are depicted using different colors. Since regulation, inhibition, catalyze, and activation are one-way relations, we model the protein functional interaction network as a directed graph. We

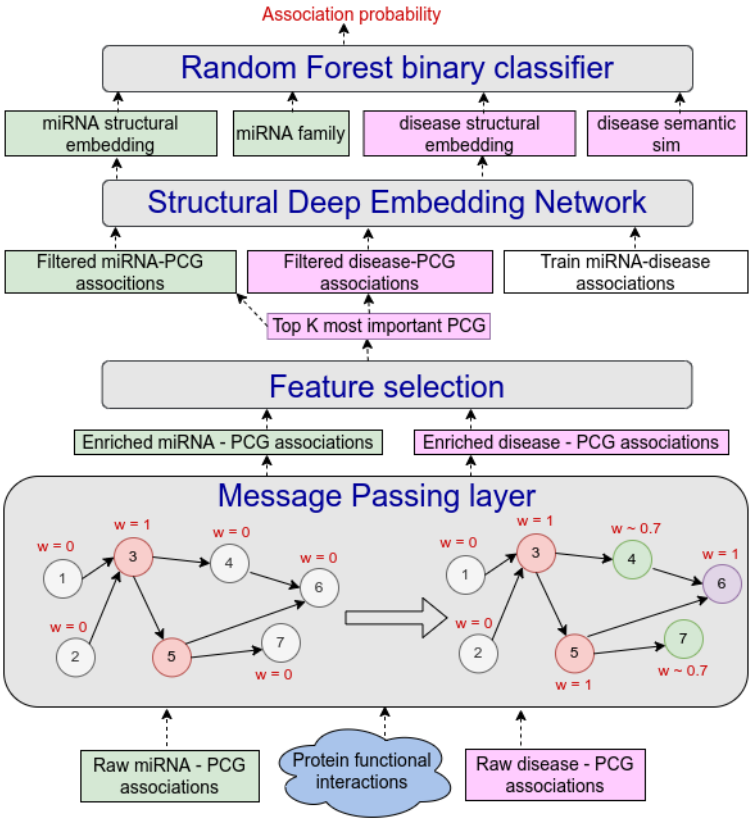


FIGURE 6.1: MPM's architecture.

retrieve the protein functional interaction network from [229] (version 2020). We generate a directed graph from the given data as follows. Each PCG is represented as a node; a protein-protein binding interaction is modeled as two directed edges. Each relation, i.e., inhibit, activate, regulate, and catalyze, is represented by a directed edge between the corresponding nodes. Overall, our protein functional interaction network consists of 423,672 directed links between 23,611 PCGs. Some PCG nodes might be isolated in the generated network because we only include experimentally verified interactions.

Modelling miRNAs using the protein functional interaction networks. We obtain the experimentally validated miRNA-PCG interactions from the miRTarBase database [86] (release 8.0 which encapsulates the gene expression profiles of more miRNAs compared to the RAIN database (employed by MUCOMID). We then model each miRNA as a network of PCGs built up from the protein functional interaction network. There is a directed link between two nodes if there is a directed link between the corresponding nodes in the functional interaction network. Each PCG node in the network has a feature vector of one dimension. The feature value of a PCG node is set to 1 if there is a known interaction between it and the current miRNA, and 0 otherwise.

Modelling diseases using the protein functional interaction networks. We obtain the disease-PCG associations from the DisGeNET [168] database, which contains one of the largest publicly available collections of genes associated

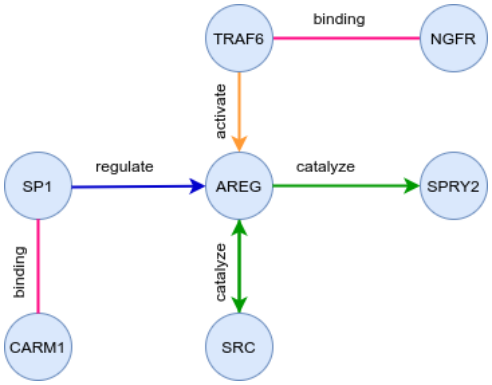


FIGURE 6.2: An example of the protein functional interaction network.

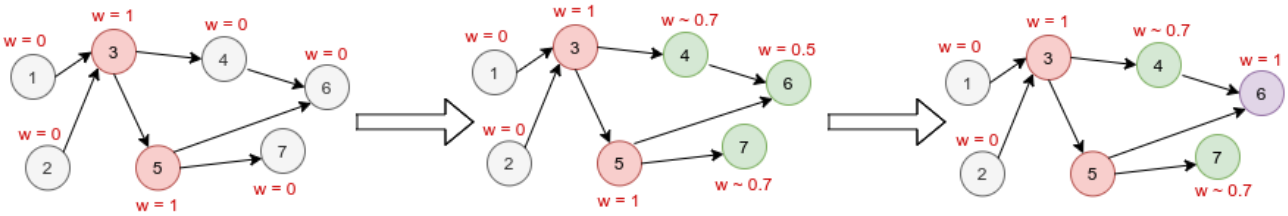


FIGURE 6.3: An example of how a message passing framework functions.

with human diseases. DisGeNET covers the information corresponding to more diseases than the DISEASES database (employed by MUCOMID). As above, we then model each disease as a network containing all PCGs from the protein functional interaction network. There is a directed link between two nodes if there is a directed link between the corresponding nodes in the functional interaction network. Each PCG in the network has a feature vector of one dimension. The feature value of a PCG node is set to be the *normalized confidence score* (in [0,1] range) of the corresponding association between the PCG and the current disease if there exists one, and 0 otherwise.

TABLE 6.1: Statistics for the side data sources. $|E|$, $|V_m|$, $|V_d|$, $|V_p|$ denote the number of interactions/associations, miRNAs, diseases, and PCGs, respectively.

NETWORK	$\|E\|$	$\|V_m\|$	$\|V_d\|$	$\|V_p\|$
MIRNA-PCG	345,357	1,618	-	23,611
DISEASE-PCG	510,782	-	3,679	23,611
PROTEIN FUNCTIONAL INTERACTIONS	423,672			23,611

The message passing framework for feature enrichment

The message passing module is responsible for further enriching the input representations via a simple message passing technique. It takes as input the miRNAs and diseases modeled using the protein functional interaction networks with the corresponding node features as described in the previous section.

MiRNA-target or disease-PCG association data might be incomplete due to the lack of biological experiments or other technical limitations. Moreover, the data acquisition methods might fail to detect *indirect* PCG associations. Our message passing strategy allows us to infer such indirect or missing miRNA-PCG and disease-PCG connections. In particular, at each iteration, a message passing step is performed in which only weights of the nodes with unknown associations (i.e., nodes with initial 0 weights) with miRNAs/diseases are updated. Formally, the inferred weight for a particular node i whose original weight is 0 at iteration t is calculated in accordance with its parents and their degrees as follows:

$$\mathbf{w}_t(i) = \frac{1}{\sqrt{\mathbf{d}_{in}(i)}} \sum_{j \in Par(i)} \frac{\mathbf{w}_{t-1}(j)}{\sqrt{\mathbf{d}_{out}(j)}} \tag{6.1}$$

where $Par(i)$ denotes the set of parent nodes of node i, $\mathbf{w}_{t-1}(j)$ is the weight of node j calculated at iteration $t-1$, $\mathbf{d}_{in}(i)$ and $\mathbf{d}_{out}(j)$ denote the in-degree and the out-degree of nodes i and j, respectively. We provide an example of how the proposed message passing layer/framework works in Figure 6.3 where the numbers inside the circles indicate nodes' IDs. 'w' indicates the node feature weight (as described in section 6.1.1). In the first iteration, new weights for nodes $4, 6, 7$ are calculated according to equation (6.1). Only the weight for node 6 gets updated during the second iteration.

The results presented in Section 6.3 correspond to the output from the message passing framework after one iteration. We choose one iteration as it acquires the best performance on all inductive test datasets.

6.1.2 The feature selection module

The disease category

The MESH ontology [15] is a well-organized vocabulary produced by the National Library of Medicine, where diseases are classified into different categories. MESH ontology can be visualized as a tree where each layer in the tree represents one level of granularity. The uppermost level represents the most general category. We obtain the disease category information from the MESH database. We assign a label to each disease that corresponds to its second-level category for "*Infection*" related diseases and its first-level category for the rest. We group all categories which have less than ten members into one common "Others" category to make the label space less sparse. In the end, each disease is assigned one of the 28 categories.

Feature selection with a side-supervised task

To remove redundant and noisy miRNA/disease-PCG associations, we employ another source of information (the disease categories as described in

Section 6.1.2) as input to our feature selection module. The rationale driving the feature selection step is that PCGs that are important for differentiating between diseases of different classes should also be indicative of the disease conditions and should, therefore, be important factors for the miRNA-disease association prediction problem.

Formally, we are given the set of diseases $\mathbf{D}$, their associated categories $\mathbf{C}$, and their inferred (up to t hop(s)) PCG association profiles $\mathbf{DI}_t$. We are interested in finding the top K most important PCG features predictive of the disease category.

As suggested in [50, 206], ReliefF [111, 161] is a competitive feature selection method for biological datasets. For that reason, we employ ReliefF to select the K most important PCGs. ReliefF estimates each feature's importance according to the relationship of n random samples to their nearest neighbors. For a given sample, the algorithm selects k nearest samples from the same class (hits) and k nearest samples from each of the other classes (misses). The feature importance is then quantified as to how well it can differentiate between the misses and the hits samples. The results presented in Section 6.3 correspond to $K = 100$ as it acquires the best performance on all inductive testing datasets.

6.1.3 The structural embedding learning

Network construction. Let $\mathbf{P}_K$ denote the set of K most informative PCGs for the disease category prediction task obtained as output from the feature selection module. Let $\mathbf{A}_p$ denote the adjacency matrix generated from the subset of PCG-PCG interactions for all PCGs in $\mathbf{P}_K$. Similarly, let $\mathbf{A}_{mp}$ be the adjacency matrix generated from the subset of miRNA-PCG associations for all PCGs in $\mathbf{P}_K$. $\mathbf{A}_{dp}$ denotes the adjacency matrix generated from the subset of disease-PCG associations for all PCGs in $\mathbf{P}_K$. Let $\mathbf{A}_{md}$ be the adjacency matrix constructed from the known miRNA-disease associations. We construct an undirected network $\mathcal{G}_{mdp}$ from the training miRNA-disease associations and the filtered sets of miRNA-PCG, disease-PCG associations, and PCG-PCG interactions. The adjacency matrix for $\mathcal{G}_{mdp}$ is then given as follows:

$$\mathbf{A}_{mdp} = \begin{bmatrix} \mathbf{Z}_m & \mathbf{A}_{md} & \mathbf{A}_{mp} \\ \mathbf{A}_{md}^T & \mathbf{Z}_d & \mathbf{A}_{dp} \\ \mathbf{A}_{mp}^T & \mathbf{A}_{dp}^T & \mathbf{A}_p \end{bmatrix}$$

where $\mathbf{Z}_m \in \mathbf{R}^{n_m \times n_m}$ and $\mathbf{Z}_d \in \mathbf{R}^{n_d \times n_d}$ are the matrices of all zeros; n_m and n_d are the number of miRNAs and diseases, respectively.

Structural Deep Network embedding. The Structural Deep Network embedding [218] is a node representation learning method that can capture the network's global and local structure efficiently by employing a deep autoencoder. The model is claimed to be able to learn highly non-linear network structures while being robust to the network sparsity [218]. In particular, SDNE enforces the first-order similarity constraint, which basically

implies that two vertices in a network are similar if they are linked by an observed edge as a supervised signal to learn the local network structure. The second-order proximity, which assumes that two vertices sharing many common neighbors are similar, is also incorporated into the model to capture the global network structure. A comparative study presented in [75] indicates that SDNE acquires the best performance compared with other structural embedding methods for the miRNA-disease association prediction problem. For that reason, we adapt SDNE to learn the structural embeddings for miRNAs and diseases from the $\mathcal{G}_{mdp}$ network. We use the SDNE implementation shared by [75] to generate the embeddings for miRNAs and diseases from the inter-connected miRNA-PCG-disease network. The results presented in Section 6.3 correspond to the SDNE with two encoder layers of size [1000, 128], one decoder layer, and the output embedding of 128 dimensions as suggested in [75].

6.1.4 The classification module

The miRNA family features. MiRNAs belonging to the same family usually share similar secondary structures and have similar biological functions [102]. Therefore, the miRNA family information is highly relevant to the miRNA-disease association prediction task. We retrieve the miRNA family information from mirBase database [113] and construct the corresponding features using the same steps as described in Section 3.2. In the end, each miRNA is assigned to one of the 1,375 families. We model each miRNA's family features as the one-hot encoding of its family.

The disease semantic similarity features. The disease semantic similarity [219, 233] quantifies how similar two particular diseases are based on their relative positions on the disease MESH ontology [15]. We use the code and the setup in [51] to compute a disease semantic similarity (ref. **DS** in Section 3.2) matrix for our 3,679 diseases set. Each entry (i,j) in the matrix indicates how similar disease i is to disease j. We model each disease's semantic similarity features as the corresponding row entry in the similarity matrix.

The classifier. The final classifier module takes the input representation for miRNA-disease pairs and for each pair, it outputs an association probability in the [0,1] range. The higher the probability, the more likely the input pair is associated. For a particular (m, d) input pair, we construct the input feature vector as the concatenation of their corresponding structural embeddings, the miRNA family, and disease semantic similarity features. More specifically, $\mathbf{X}_{md} = [\mathbf{E}_m, \mathbf{E}_d, \mathbf{F}_m, \mathbf{S}_d]$, where X_{md} denotes the input feature vector corresponding to (m, d); $\mathbf{E}_m, \mathbf{E}_d$ represent the pre-trained embeddings output from SDNE; while $\mathbf{F}_m$ refers to the miRNA family feature for miRNA m; $\mathbf{S}_d$ corresponds to the disease semantic similarity for disease d. A pictorial illustration of the final miRNA-disease pair representation is given in Figure 6.4. We train a Random Forest classifier[17, 193] with 350 estimators to do the association prediction task.

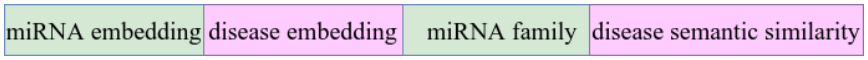


FIGURE 6.4: The final miRNA-disease input pair representation.

6.2 The Experimental data and setup

6.2.1 Compared models

We compare our model with six recently proposed methods: (i) EPMDA [55], DBMDA [244], and NIMGCN [126], which utilize hand-crafted features derived from known miRNA-disease associations, (ii) MUCOMID [52] and DIMIG 2.0 [162], which use graph convolution networks (GCNs) for feature extraction from various interaction networks (iii) NEMII [75] which employs hand-crafted features as well as the latent features extracted using a graph embedding method. A detailed description of benchmarked models is presented in Section 3.3. As an ablation study, we compare MPM with four of its simpler variants as summarized in Table 6.8.

6.2.2 The miRNA-disease association data source

We retrieve the set of miRNA-disease associations from the HMDD v2.0 [129] and HMDD v3.0 [88] databases. We then perform various preprocessing and filtering steps as described in the "Data acquisition and preprocessing" section. In the end, the filtered data for the HMDD v2.0 database (denoted as HMDD2) contains 4,592 known associations between 442 miRNAs and 309 diseases. The filtered data for the HMDD v3.0 database (referred to as HMDD3) includes 10,494 known associations between 742 miRNAs and 545 diseases. Note that the two datasets presented here also differ from those in chapters 4 and 5 because of the difference in preprocessing steps.

Data acquisition and preprocessing

As the quantity and quality of the employed data source greatly impact the predictive power of the learned models, apart from the model development, our contribution also lies in the data acquisition and preprocessing. In the following sections, we describe our data acquisition and preprocessing steps.

Disease ID matching. The data deposited in HMDD 2.0 and HMDD 3.0 only provides disease names. Even worse, different names might refer to the same diseases. In addition, to retrieve the disease ontology or disease-PCG associations, we need the diseases' MESH IDs. Therefore, in the first steps of our preprocessing pipeline, we match the HMDD 2.0's and HMDD 3.0's disease names with their corresponding MESH IDs.

In order to do that, we first collect the list of disease IDs, along with their names and synonyms, from the MESH database [15]. We then standardize all disease names and synonyms (remove redundant spaces and quotations, and convert all to lowercase). After that, our disease matcher works as follows: (i) if there is an exact match between the searching disease name and any

MESH names/synonyms, then it assigns the corresponding MESH ID to that disease name (ii) otherwise, it outputs a list of names along with their MESH IDs which are the most similar to the searching name and only contain up to several different characters in the character sequence. We later quickly reviewed these lists to increase the data coverage as much as possible.

miRNA name standardization. The HMDD 2.0 and HMDD 3.0 databases store the known associations reported in scientific publications and do not reflect the changes in the miRNA knowledgebase over time. Therefore, the same miRNA might appear with different IDs in the miRNA-disease association databases. To remove unnecessary noise and make the data consistent, we standardize the miRNA IDs according to miRBase [113] - one of the most reliable and popular databases to retrieve miRNAs related information. More specifically, we match multiple miRNAs aliases together and obsoleted IDs to the newly assigned ones according to the data retrieved from miRBase, version 22.1. Table 6.2 presents the statistics associated with the number of miRNAs and miRNA-disease associations after standardization.

TABLE 6.2: Before and after miRNA name standardization data statistics. n_m and n_{md} refer to the number of miRNAs and miRNA-disease associations, respectively.

Database	Before		After	
	n_m	n_{md}	n_m	n_{md}
HMDD 2.0	578	6,401	540	5,909
HMDD 3.0	1,120	15,165	859	12,552

6.2.3 The data set up

While the K-fold cross-validation (K-fold CV) technique is widely used among existing works, it is insufficient to evaluate the models' performance on completely new diseases, given the small size of the association datasets. Therefore, besides 5-fold CV evaluation on the HMDD2 and HMDD3 datasets, we here propose and employ two realistic testing setups: *transductive* and *inductive* to evaluate and compare models. The transductive testing setup aims at evaluating different models' performances on a larger, independent test set which contains the newly discovered associations between the miRNAs and diseases that have already been observed with some previously known associations during the training phase. In this setup, we train each model on the HMDD2 dataset and test it on the HELD-OUT1 test set. HELD-OUT1 contains only associations corresponding to the miRNAs and diseases that are observed in the HMDD2 dataset. However, the known associations in HELD-OUT1 do not appear in the training set HMDD2. The inductive testing setup aims at evaluating models' performance on completely new diseases and new miRNAs. In this setup, we conduct large-scale experiments on the 20 independent test sets to test each model's performance on (i) a dataset with many new miRNAs (the NOVEL-MIRNA test set), (ii) 18 complete test sets for new diseases, and (iii) a dataset with many new miRNAs and new diseases

(the HELD-OUT2 test set). For the evaluation with the NOVEL-MIRNA and HELD-OUT2 test sets, we train the benchmarked models with the HMDD2 dataset. For the evaluation related to 18 new diseases, we train all models with all available association data for any disease other than the ones in the test sets. All datasets' statistics are presented in Table 6.3 and Table 6.4.

In the following, we present details regarding our generated datasets.

The transductive testing setup

The transductive testing setup aims at evaluating different models' performances on the set of partially observed miRNAs and diseases. We train each model with the HMDD2 dataset while testing them with the HELD-OUT1 test set as described below.

Let $\mathbf{M}$ and $\mathbf{D}$ denote the set of miRNAs and diseases observed in the HMDD2 dataset, correspondingly. We construct the HELD-OUT1 dataset by restricting the set of miRNAs and diseases to $\mathbf{M}$ and $\mathbf{D}$ and including only the miRNA-diseases associations, which appear in the HMDD3 dataset but not in the HMDD2 dataset. A mathematical description of HELD-OUT1 is given below:

$$\text{HELD-OUT1} = (\mathbf{M} \times \mathbf{D} \cap \text{HMDD3}) \backslash \text{HMDD2}$$

Where $\mathbf{M} \times \mathbf{D}$ denotes the set of all possible pair combinations between miRNAs in $\mathbf{M}$ and diseases in $\mathbf{D}$. Table 6.3 presents the transductive training and testing data statistics. We generate the negative training and testing samples using the negative sampling strategy given in section 6.2.4.

The inductive setting setup

The large independent testing sets

The HELD-OUT2 *test set.* HELD-OUT2 contains all associations that appear in HMDD3 but not in HMDD2. We devise this dataset to test all models' performance on a large independent test set that contains both new miRNAs and new diseases (with respect to the training data). After preprocessing, HELD-OUT2 contains 6,388 known associations for 697 miRNAs and 509 diseases. Among those, there are 300 new miRNAs and 282 new diseases that do not appear in the training set HMDD2.

The NOVEL-MIRNA *test set.* The NOVEL-MIRNA test set is a subset of the HELD-OUT2 test set. To construct NOVEL-MIRNA, we remove all associations related to any disease that does not appear in $\mathbf{D}$. In the end, NOVEL-MIRNA contains 4,734 known associations for 638 miRNAs and 227 diseases in which there are 256 new miRNAs that do not appear in the training set HMDD2. The data statistics for our large independent test sets are presented in Table 6.3. Note that the datasets presented here differ from those in chapter 5 because of the difference in data preprocessing steps.

The datasets for new diseases

TABLE 6.3: The association data statistics. $|n_{md}|$, $|n_m|$, $|n_d|$ refer to the number of associations, miRNAs and diseases, respectively.

DATASET	$\|n_{md}\|$	$\|n_m\|$	$\|n_d\|$
HMDD2	4,592	442	309
HMDD3	10,494	742	545
HMDD2 ∪ HMDD3	10,980	742	591
HELD-OUT1	4,311	382	226
HELD-OUT2	6,388	697	509
NOVEL-MIRNA	4,734	638	227

TABLE 6.4: Statistics for the datasets corresponding to new diseases. $|E_{train}|$ and $|E_{test}|$ refer to the number of *positive* training and testing samples, respectively.

DISEASE	$\|E_{train}\|$	$\|E_{test}\|$	DISEASE	$\|E_{train}\|$	$\|E_{test}\|$
D001943	10649	331	D005909	10803	177
D015179	10704	276	D001749	10813	167
D013274	10720	260	D012516	10821	159
D008175	10757	223	D010190	10822	158
D011471	10761	219	D006333	10838	142
D002289	10766	214	D002292	10839	141
D010051	10789	191	D003110	10854	126
D008545	10791	189	D015470	10868	112
D005910	10791	189	D002294	10876	104

The inductive setting setup aims at evaluating models' performances on completely new diseases and is described as follows:

- Let H = HMDD2 ∪ HMDD3
- We take out the set of diseases $\hat{\mathcal{D}}$ such that each disease $d \in \hat{\mathcal{D}}$ has more than 100 known associations in H. There are 18 such diseases.
- For each disease $d \in \hat{\mathcal{D}}$, a dataset is created as follows: (i) The positive training set includes all known associations in H except those associated with d, (ii) The negative training samples are generated according to section 6.2.4, (iii) We evaluate all models on the ***complete*** testing set where all known associations for d in H form the positive test set and the negative testing samples consist of all possible combinations of d and any miRNA that does not appear in the positive testing set.

Table 6.4 presents the statistics corresponding to the 18 datasets for new diseases.

6.2.4 The negative sampling strategy.

We define the negative pool as the set of all possible combinations of miRNA-disease pairs that do not appear in the set of all known associations. For all training data, we fix the negative:positive ratio to 1:1. For the independent testing sets (HELD-OUT1, NOVEL-MIRNA, and HELD-OUT2), we vary the

ratio to be one of [1:1, 1:5, 1:10]. For each negative:positive sample rate, we randomly draw 10 subsets from the negative pool and evaluate all models' performance on all those sampled sets to avoid bias and make the comparison as fair as possible. In summary, in the transductive setting, we have 10 train and 10 test sets (corresponding to different negative sample sets). We evaluate each model by training it on all 100 train and test set combinations, each with 2 random model initialization. In total, we report the average results corresponding to 200 experimental runs for the transductive setting.

We use the entire set of unknown interactions as negative test samples for the experiments corresponding to the new disease datasets. For each dataset, we run the model with 10 sampled train sets. For each set, we run the model twice with different random initializations. The reported results presented in Section 6.3 are the average results over 20 experimental runs.

6.2.5 Evaluation Metrics.

For non-parametric metrics, we report the Area under the Receiver Operating Characteristic (AUC), the Average Precision (AP) (which summarizes the Precision-Recall curve). For threshold-based metrics, we report the Sensitivity (or Recall (Rec), referred to as SN), Specificity (SP), Accuracy (ACC), Precision (Pre), F1, and Matthews correlation coefficient (MCC) scores. Besides, for the new disease test sets, we also report the number of correctly predicted miRNA-disease associations among the top 100 highest predicted scores (denoted as Top100) generated by the benchmarked models. For all tables, bold font is used to highlight the best scores.

6.2.6 Hyperparameter setup and implementation details

MPM and its variants. We experiment with the number of message passing iteration t in [1, 2, 10]. For the feature selection module, we run ReliefF [111] with 20 neighbors and the number of selected features K from 50 to 500 with a step size of 50. The results reported in Section 6.3 correspond to $t = 1$, and $K = 100$, which result in the best average AP score among 18 datasets in the inductive test setting. For SDNE, we use the default parameter as suggested by NEMII [75] with the embedding size fixed to 128. The Random Forest classifier is trained with 350 estimators.

Existing benchmarked models. For EPMDA, DBMDA, and NIMGCN, we use the code and setup released in [51]. For NEMII and MUCOMID, we use the same code and setup as published by the authors. For DIMIG 2.0, we follow the same testing strategies employed in [52].

6.3 Results

6.3.1 MPM vs. existing works (SOTA)

Tables 6.5 and 6.6 present the average performance scores for all benchmarked models on our 21 large test sets in the transductive and inductive testing setups. In Table 6.5, we report the average AP and AUC scores corresponding to different positive:negative testing sample rates. We do not have the results for EPMDA on the 18 test sets for new diseases because all pairs' representations are zeros since new diseases appear as isolated nodes in the network for the topology-based feature extraction. Table 6.7 shows the results corresponding to the 5-fold CV results on the HMDD2 and HMDD3 datasets. For each dataset, we randomly split the data according to 5 different random seeds and report the average performance.

TABLE 6.5: Results on the three large test sets. *nr* denotes the positive:negative sample rate.

dataset	*nr*	NIMGCN		DBMDA		EPMDA		NEMII		MUCOMID		DIMIG 2.0		MPM		SOTA↑
		AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AP
HELD	1:1	.542	.554	.657	.622	.698	.624	.838	.831	.832	.826	.499	.5	**.848**	**.844**	1.6%
-OUT1	1:5	.541	.207	.656	.256	.698	.256	.838	.542	.832	.534	.499	.167	**.848**	**.573**	5.7%
	1:10	.542	.118	.656	.149	.698	.148	.838	.395	.832	.385	.499	.091	**.848**	**.429**	8.6%
NOVEL	1:1	.532	.549	.644	.621	.716	.643	.865	.857	.827	.819	.499	.5	**.869**	**.866**	1.1%
-MIRNA	1:5	.53	.202	.645	.261	.718	.281	.866	.597	.827	.519	.499	.167	**.87**	**.62**	3.4%
	1:10	.53	.115	.645	.153	.719	.167	.866	.452	.827	.37	.499	.091	**.87**	**.479**	6.0%
HELD	1:1	.513	.517	.638	.617	.704	.648	.859	.853	.811	.812	.499	.5	**.863**	**.865**	1.4%
-OUT2	1:5	.513	.176	.638	.257	.703	.291	.859	.581	.812	.514	.499	.167	**.863**	**.621**	6.9%
	1:10	.512	.097	.638	.015	.704	.176	.858	.435	.811	.368	.499	.091	**.862**	**.485**	11.5%

TABLE 6.6: Results on the 18 inductive testing sets for new diseases.

dataset	Method	AUC	AP	SN	SP	Acc	Pre	F1	Mcc	Top100
D001943	MPM	.895	.824	55.7	98.6	**89.8**	**91.4**	**69.1**	**66.3**	**99.4**
	NIMGCN	.344	.16	0	**99.4**	79.1	0	0	-2.3	.0
	DBMDA	.773	.507	**62.4**	84.8	80.2	56	57.9	46.3	67.1
	NEMII	**.916**	**.827**	52.5	98.6	89.2	91.1	66.3	63.9	98.4
	MUCOMID	.669	.414	42.8	75.2	68.6	42.8	36.1	21.1	52.8
	DIMIG 2.0	.5	.205	**100**	0.0	20.5	20.5	34.0	0.0	20.0
D015179	MPM	.905	**.806**	74.2	95.1	**91.5**	**75.6**	**74.9**	**69.8**	**95.2**
	NIMGCN	.373	.134	0	**98.9**	82	0	0	-2.6	.0
	DBMDA	.792	.463	**77.1**	73.5	74.1	44.3	54.8	43.3	53.8
	NEMII	**.910**	.797	73.4	94.5	90.9	73.4	73.4	67.9	94.7
	MUCOMID	.651	.34	76.1	49.5	54	25.8	37.6	20.3	41.3
	DIMIG 2.0	.499	.171	100	0	17.1	17.1	29.1	0	11.5
D013274	MPM	.938	**.837**	82	92.7	91	**68.3**	74.5	69.5	**96.0**
	NIMGCN	.394	.132	0	**99.9**	83.9	0	0	-.8	.0
	DBMDA	.872	.503	**84**	84.2	**84.2**	50.9	63.3	56.8	54.5

	NEMII	**.952**	.835	83.8	92.1	90.8	67.2	**74.5**	**69.7**	94.5
	MuCoMiD	.597	.249	69.8	46.7	50.4	21	31.6	12.9	28.8
	DimiG 2.0	.498	.161	100	0	16.1	16.1	27.7	0	18.5
	MPM	.926	**.764**	**92.3**	76.5	78.7	39.4	55	51	**87.6**
	Nimgcn	.374	.115	0	**97.9**	84.4	0	0	-4.1	.0
	Dbmda	.829	.437	74.2	84	82.7	44.7	55.2	48.2	52.4
D008175	NEMII	**.935**	.749	88.5	84.9	**85.4**	**48.4**	**62.6**	**58.3**	85.3
	MuCoMiD	.734	.375	55.6	72.4	70.1	37.3	36.7	27.4	43.4
	DimiG 2.0	.499	.138	100	0	13.8	13.8	24.2	0	13.0
	MPM	.922	**.733**	80.1	91.5	**90**	59.6	68.3	63.5	**85.5**
	Nimgcn	.404	.116	0	**99.1**	85.7	0	0	-2.1	.0
	Dbmda	.747	.395	64.8	75	73.6	40.4	47.4	35.6	52.6
D011471	NEMII	**.926**	.653	**78.2**	92.2	90.3	**61.1**	**68.6**	**63.6**	72.2
	MuCoMiD	.692	.33	42.3	71.2	67.3	38.8	21.8	15.2	39.1
	DimiG 2.0	.499	.135	100	0	13.5	13.5	23.8	0	14.5
	MPM	.934	**.802**	40	**99.5**	91.6	**92.4**	55.7	57.5	**92.1**
	Nimgcn	.385	.108	0	99.3	86.2	0	0	-2.1	.0
	Dbmda	.661	.303	**62.9**	64.1	64	33	39.7	24.9	39.8
D002289	NEMII	**.947**	.800	59.9	98.0	**92.9**	81.9	**69.0**	**66.2**	90.1
	MuCoMiD	.676	.278	61.1	55.9	56.6	28.3	27.7	16.6	31.5
	DimiG 2.0	.498	.132	**100**	0.0	13.2	13.2	23.4	0.0	10.5
	MPM	.958	**.792**	**87.9**	92.5	91.9	61.1	**72.0**	**69.0**	**87.1**
	Nimgcn	.478	.114	0	**99.9**	88.1	0	0	-.6	.0
	Dbmda	.845	.4	81.4	79	79.3	40.1	52.9	46.8	46.3
D010051	NEMII	**.960**	.76	54.7	97.9	**92.8**	**79.6**	63.7	61.9	84.7
	MuCoMiD	.774	.388	81.4	59	61.6	24.1	35.8	28.3	45.5
	DimiG 2.0	.498	.118	**100**	0.0	11.8	11.8	21.1	0.0	10.5
	MPM	.917	**.724**	52.1	98	**92.6**	77.6	**62.1**	**59.8**	**82.5**
	Nimgcn	.466	.108	0	**99.9**	88.3	0	0	-.6	.0
	Dbmda	.766	.355	70.4	69.9	69.9	33.4	42.4	32.9	44.6
D008545	NEMII	**.928**	.706	31.4	99.3	91.4	**86.1**	45.3	48.4	79.5
	MuCoMiD	.755	.365	**52.5**	72.2	69.9	38.7	31.4	25.6	43.4
	DimiG 2.0	.499	.117	100	0	11.7	11.7	20.9	0	8.0
	MPM	.947	**.759**	**49**	98.4	**92.6**	80.9	59.3	**58.6**	**81.6**
	Nimgcn	.489	.112	0	**99.9**	88.3	0	0	-.6	.0
	Dbmda	.644	.246	36.5	83.9	78.4	30.5	31.9	20.7	32.8
D005910	NEMII	**.953**	.731	50.7	98.1	92.6	**77.8**	**60.0**	58.5	80.1
	MuCoMiD	.786	.409	32.2	88.8	82.2	46	28.3	24.6	48.0
	DimiG 2.0	.499	.117	**100**	0.0	11.7	11.7	20.9	0.0	10.0
	MPM	.938	**.736**	75.1	95.6	**93.3**	**67.6**	**71.1**	**67.5**	79.7
	Nimgcn	.563	.123	0	**99.9**	89	0	0	-.8	.0
	Dbmda	.814	.369	82.0	67.9	69.4	30.7	42.9	36.4	42.5
D005909	NEMII	**.943**	.712	**86.7**	89.1	88.8	49.4	62.9	60.0	78.1
	MuCoMiD	.815	.418	55.4	84.9	81.7	38.3	42.1	35.3	46.5
	DimiG 2.0	.499	.109	100	0	10.9	10.9	19.7	0	12.0
	MPM	.94	**.785**	**88.3**	84.7	85.1	40.1	55.1	53	**88.7**
	Nimgcn	.432	.089	0	**99.9**	**89.6**	0	0	-.9	.0
	Dbmda	.815	.340	76.3	75.6	75.7	31.7	43.8	37.9	40.2
D001749	NEMII	**.948**	.767	80.8	91.4	90.3	**52.1**	**63.3**	**59.9**	86.4
	MuCoMiD	.808	.446	43.1	90.7	85.7	52	39	36.4	52.1

	DIMIG 2.0	.498	.103	100	0	10.3	10.3	18.7	0	12.5
D012516	MPM	.932	**.713**	44	99	**93.6**	**82.6**	**57.3**	**57.4**	**78.8**
	NIMGCN	.821	.262	0	**99.9**	90.1	0	0	-.5	6.5
	DBMDA	.755	.323	55.6	87.7	84.6	36.9	43.6	36.8	47.6
	NEMII	**.936**	.658	42.3	98.3	92.8	73.5	53.2	52.2	73.4
	MUCOMID	.781	.349	61.3	79.5	77.7	30.5	38.2	31.6	40.7
	DIMIG 2.0	.499	.098	100	0	9.8	9.8	17.9	0	13.5
D010190	MPM	.944	**.749**	83.1	94.2	93.1	60.9	70.2	**67.5**	**79.7**
	NIMGCN	.448	.088	0	**99.8**	90.1	0	0	-1	.0
	DBMDA	.871	.366	**84.4**	82	82.2	33.8	48.3	46	39.6
	NEMII	**.947**	.744	71.7	96	**93.6**	**66.2**	**68.5**	65.2	76.7
	MUCOMID	.784	.373	18.6	96.8	89.2	52.1	23.7	23.5	44.7
	DIMIG 2.0	.499	.098	100	0	9.8	9.8	17.8	0	9.0
D006333	MPM	.949	**.669**	**67.2**	95.9	**93.4**	**61.2**	**64.0**	60.5	**68.8**
	NIMGCN	.729	.18	0	**99.9**	91.2	0	0	-.5	4.2
	DBMDA	.743	.299	58.5	85.9	83.5	32.5	41	35.1	44.1
	NEMII	**.953**	.651	60	96.2	93	60.2	60	**56.2**	65.9
	MUCOMID	.816	.395	36.5	88.3	83.8	46	29.3	28.1	45.8
	DIMIG 2.0	.499	.088	100	0	8.8	8.8	16.1	0	6.0
D002292	MPM	.939	**.684**	80.0	93.1	91.9	52.4	**63.3**	**60.7**	**70.8**
	NIMGCN	.471	.082	0	**99.9**	91.2	0	0	-.6	.0
	DBMDA	.739	.238	52.5	84.4	81.6	27.8	35.6	28.7	32.5
	NEMII	**.945**	.653	74.3	93.6	**91.9**	**52.8**	61.6	58.3	67.0
	MUCOMID	.774	.285	**86.5**	53.2	56.1	16.7	27.4	23.8	31.8
	DIMIG 2.0	.498	.087	100	0	8.7	8.7	16	0	5.5
D003110	MPM	**.945**	**.659**	**99.2**	19.6	25.8	9.7	17.7	12.6	**68.3**
	NIMGCN	.441	.069	0	**99.6**	**91.8**	0	0	-1.2	.0
	DBMDA	.823	.242	91.3	70.4	72.1	**23.5**	36.8	37.2	23.8
	NEMII	0.938	0.600	91.0	82.5	83.2	**30.7**	**45.9**	**46.8**	60.8
	MUCOMID	.764	.271	97.4	13.2	19.8	8.8	16.1	7.5	31.0
	DIMIG 2.0	.498	.078	100	0	7.8	7.8	14.5	0	7.5
D015470	MPM	**.953**	**.655**	70.9	95.3	93.6	52.8	60.5	**57.8**	**63.4**
	NIMGCN	.732	.158	0	**99.9**	93	0	0	-.6	3.9
	DBMDA	.737	.259	**64**	74.2	73.5	23.8	32.8	27.4	36.1
	NEMII	.951	.625	53.4	97.4	**94.4**	**61.6**	56.7	54.2	60.6
	MUCOMID	.779	.29	46.5	82.9	80.4	24.2	27.3	22.8	29.6
	DIMIG 2.0	.498	.069	100	0	6.9	6.9	13	0	8.5
D002294	MPM	**.962**	**.669**	90.6	92.5	92.4	**45.4**	60.5	60.9	**62.8**
	NIMGCN	.834	.186	0	**99.9**	**93.5**	0	0	-.6	.55
	DBMDA	.789	.241	**68.6**	80.8	80	22.8	33.3	31.3	31.8
	NEMII	.956	.608	90.1	90.8	90.8	40.4	**55.7**	**56.6**	59.5
	MUCOMID	.855	.384	90.9	62.5	64.3	19.7	30.2	31.1	38.7
	DIMIG 2.0	.498	.064	100	0	6.4	6.4	12.1	0	8.5

In the three large independent test sets (ref. Table 6.5), MPM outperforms all benchmarked models (SOTA) on the HELD-OUT1 (transductive setting), NOVEL-MIRNA (with many new miRNAs), and HELD-OUT2 (with new miRNAs and new diseases) test sets with a gain of up to 11.5% in AP score. The gains are more significant when more negative samples are added

TABLE 6.7: Results for 5-fold cross-validation on the HMDD2 and HMDD3 datasets.

dataset	Method	AUC	AP	SN	SP	ACC	Pre	F1	MCC
HMDD2	MPM	0.89	**0.9**	80.7	81.5	81.1	81.3	81.0	62.2
	NIMGCN	0.88	0.87	70.2	**84.2**	77.2	77.9	71.0	54.6
	DBMDA	0.72	0.68	66.9	72.4	69.7	70.8	68.8	39.4
	EPMDA	0.52	0.61	36.0	64.0	50.0	18.0	24.0	0.0
	NEMII	0.9	**0.9**	81.4	81.5	81.4	81.5	81.4	62.9
	MUCOMID	**0.91**	**0.9**	83.0	82.5	**82.8**	**82.7**	**82.8**	**65.6**
	DIMIG 2.0	0.5	0.51	**100.0**	0.0	50.0	50.0	66.7	0.0
HMDD3	MPM	0.91	0.91	83.8	82.0	82.9	82.3	83.0	65.8
	NIMGCN	0.89	0.89	84.6	80.7	82.7	81.5	83.0	65.4
	DBMDA	0.76	0.71	71.6	74.4	73.0	73.7	72.6	46.1
	EPMDA	0.48	0.59	48.0	52.0	50.0	24.0	32.0	0.0
	NEMII	0.91	0.91	84.1	82.0	83.0	82.4	83.2	66.1
	MUCOMID	**0.92**	**0.92**	85.2	**84.0**	**84.6**	**84.2**	**84.7**	**69.2**
	DIMIG 2.0	0.5	0.5	**100.0**	0.0	50.0	50.0	66.7	0.0

to the testing data. On the complete test sets for new diseases, MPM consistently acquires the highest Top100 scores in all test sets. Besides, MPM gains the highest AP scores in 17 out of 18 datasets. In the 5-fold CV evaluation setup, MUCOMID gains the highest performance in most reported metrics. MPM closely follows NEMII with slightly worse performance. Nonetheless, compared to the best-performing model (MUCOMID), MPM attains an equal AP score in the HMDD2 dataset and a 0.01 lower AP score in the HMDD3 dataset.

In both transductive and inductive testing setups, we observe similar trends with large performance gaps among the state-of-the-art methods. In the three large independent test sets (HELD-OUT1, NOVEL-MIRNA, HELD-OUT2), DIMIG 2.0 performs the worst, followed by NIMGCN, then DBMDA, EPMDA, MUCOMID, and then NEMII. In the 18 complete test sets for new diseases, regarding the AP scores, the order is slightly changed to NIMGCN, followed by DIMIG 2.0, then DBMDA, MUCOMID, and then NEMII. DIMIG 2.0 is a recently proposed model that formulates the miRNA-disease association prediction problem as a semi-supervised node classification task with diseases as labels. The model can integrate information from four additional knowledge sources (miRNA-PCG, disease-PCG associations, PCG-PCG interactions, and disease ontology) but only performs training using the known disease-PCG association set. Though DIMIG 2.0 can generate predictions for new miRNAs and new diseases, the large and sparse label set and the weak training signals lead to its limited predictive performance. With all AUC scores close to 0.5, the model does not perform better than a random guess.

NIMGCN performs the worst compared to other supervised baselines because it only relies on the miRNA functional and disease semantic similarities to construct the networks for the feature learning. The miRNA functional similarity is heavily biased toward well-known diseases and cannot generalize well to new diseases [219]. Also, new miRNAs appear as isolated nodes in the network and will get completely random representations. Therefore,

NIMGCN's prediction capability is limited for the little-known or completely new miRNAs or diseases.

Regarding the input sources, DBMDA improves over NIMGCN by integrating another biologically-related information source: the miRNA sequence similarity. DBMDA gains significantly better performance than NIMGCN but is still much lower than MUCOMID, NEMII, and MPM in most test sets, suggesting that the miRNA sequence similarity does bring additional benefit, but the gains are not too significant.

EPMDA proposes a topologically related feature extraction technique for miRNA-disease pair representation. Unlike most existing works, which focus on learning effective representations for miRNAs and diseases separately, EPMDA learns the miRNA-disease pair representation directly as a property of the miRNA-disease heterogeneous network constructed from the miRNA and disease Gaussian Interaction Profile kernel similarities and the miRNA-disease known associations. Even though EPMDA does not employ any additional information sources, its performance is still better than NIMGCN and DBMDA. This suggests that learning the pair representation directly from the heterogeneous network with raw miRNA-disease associations is a fruitful direction. Nonetheless, the edge perturbation score has at least $O(n^3)$ time complexity and cannot scale well to a large network [51]. Besides, fine-tuning the network cycle length parameter is not a trivial task [51].

MUCOMID proposes a multitask learning model that integrates five additional information sources to overcome the data scarcity problem. Though promising, the model applies hard-threshold filtering to filter out redundant information in the additional information sources. The results reported in Tables 6.5 and 6.6 correspond to MUCOMID's performance without the filtering step (since not all of our data have the interaction/association confidence scores available). The thresholds need to be fine-tuned for each dataset separately. For that reason, it requires considerable time and effort for parameter fine-tuning in order to employ MUCOMID for a completely new dataset. This points to an important aspect of information integration which focuses on effectively controlling/managing the quality and quantity of the added knowledge sources. Nonetheless, MUCOMID gains the highest performance in the 5-fold CV testing setup. Also, the method shows promising performance, which overcomes the problems associated with hand-crafted similarity-based methods in all testing setups.

NEMII learns structural embeddings directly from the miRNA-disease bipartite network constructed from the known miRNA-disease association data. Besides, the model is further informed by information from the miRNA family and disease semantic similarity. Though new miRNAs and new diseases get completely random representation from the structural embedding learning module, NEMII's performance on the 20 inductive testing datasets is still one of the highest, thanks to the biological information from the miRNA family and disease semantic similarity features. Overall, the effective feature extraction strategy, combined with the domain knowledge from the added side information sources, helped NEMII gain the highest performance scores

among state-of-the-art methods on most testing datasets. These results support the exploitation of structural information from the miRNA-disease association data and the importance of information integration.

MPM improves over state-of-the-art methods with a parameter-free yet effective mechanism to control the quality and quantity of the added information sources. At the same time, it addresses the existing limitation in the NEMII model by integrating additional biological relations to the new miRNAs and new diseases. The learned signals from the well-studied miRNAs/diseases will be transferred to the diseases (with only scarce knowledge) via their associated PCGs. These improvements help MPM gain state-of-the-art performance on 20 out of the 21 independent test sets in both transductive and inductive testing setups with a gain of up to 11.5% in AP score.

6.4 Ablation studies

6.4.1 MPM and simpler variants

Here, we compare MPM with four of its simpler variants as summarized in Table 6.8. MPM-NO-MP is a variant of MPM without the message passing layer that takes the raw miRNA-PCG and disease-PCG associations as input to the feature selection and structural embedding learning modules. Similarly, MPM-NO-FS is a variant of MPM without the feature selection module. The structural embedding learning module encapsulates all enriched miRNA-PCG and disease-PCG associations output from the message passing layer into its heterogeneous network for learning node embeddings. MPM-NO-MPFS is a variant of MPM without the message passing and the feature selection modules. The heterogeneous network input to SDNE simply integrate all raw miRNA-PCG, disease-PCG associations retrieved from miRTarBase[86] and DisGeNET [168]. MPM-NO-SDNE is a variant of MPM in which there is no structural embedding learning. Instead, the pair representation for a particular miRNA-disease pair is the concatenation of the enriched and filtered miRNA-PCG, disease-PCG associations, miRNA family, and disease semantic similarity features.

TABLE 6.8: Simpler variants of MPM.

Model	Message Passing	Feature Selection	SDNE	PCG associations
MPM-NO-MP	X	✓	✓	✓
MPM-NO-FS	✓	X	✓	✓
MPM-NO-SDNE	✓	✓	X	✓
NEMII [75]	X	X	✓	X
MPM-NO-MPFS	X	X	✓	✓

Table 6.9 presents the results for MPM and its variants on three large independent test sets. Table 6.10 reports the results for the 18 inductive testing datasets for new diseases. We observe that MPM supersedes all of its simpler variants on the transductive testing set (HELD-OUT1), two inductive testing sets with many new miRNAs (NOVEL-MIRNA and HELD-OUT2), and 15 out

TABLE 6.9: Results for MPM and its simpler variants on the three large test sets. *nr* denotes the positive:negative sample rate.

dataset	*nr*	MPM-NO-MP		MPM-NO-FS		MPM-NO -MPFS		MPM-NO -SDNE		MPM	
		AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
HELD -OUT1	1:1	.846	.84	.814	.809	.824	.816	.837	.83	**.848**	**.844**
	1:5	.846	.564	.814	.503	.824	.516	.837	.546	**.848**	**.573**
	1:10	.847	.418	.814	.357	.824	.369	.837	.401	**.848**	**.429**
NOVEL -MIRNA	1:1	.866	.859	.823	.818	.836	.828	.842	.834	**.869**	**.866**
	1:5	.866	.602	.823	.519	.836	.538	.842	.552	**.87**	**.62**
	1:10	.867	.46	.823	.519	.823	.373	.836	.391	**.87**	**.479**
HELD -OUT2	1:1	.859	.86	.814	.819	.831	.832	.846	.847	**.863**	**.865**
	1:5	.859	.607	.814	.533	.831	.554	.846	.581	**.863**	**.621**
	1:10	.859	.468	.814	.391	.831	.411	.846	.439	**.862**	**.485**

TABLE 6.10: AP scores of MPM and its variants on 18 test sets for new diseases.

Disease	MPM	MPM-NO -MP	MPM-NO -FS	MPM-NO -MPFS	MPM-NO -SDNE
D001749	**0.785**	0.77	0.567	0.589	0.58
D001943	**0.824**	0.811	0.679	0.693	0.654
D002289	**0.802**	0.795	0.662	0.678	0.589
D002292	**0.684**	0.67	0.51	0.525	0.531
D002294	**0.669**	0.646	0.529	0.531	0.493
D003110	**0.659**	0.619	0.487	0.54	0.515
D005909	**0.736**	0.726	0.597	0.63	0.523
D005910	0.759	**0.767**	0.642	0.66	0.626
D006333	0.669	**0.671**	0.578	0.602	0.566
D008175	**0.764**	0.751	0.615	0.62	0.611
D008545	**0.724**	0.715	0.58	0.598	0.558
D010051	**0.792**	0.782	0.505	0.654	0.579
D010190	0.749	**0.761**	0.589	0.622	0.598
D011471	0.733	**0.738**	0.618	0.633	0.569
D012516	**0.713**	0.699	0.546	0.585	0.55
D013274	**0.837**	0.811	0.657	0.693	0.643
D015179	**0.806**	0.785	0.645	0.693	0.614
D015470	**0.655**	0.653	0.509	0.513	0.497

of 18 complete test sets for new diseases. The gains are the most significant on the three independent test sets (c.f. Table 6.9), especially when more negative testing samples are added. These results support the contribution of each added component. At the same time, they validate our choice of architecture.

Besides, among the simpler variants, we observe a considerable performance drop on the variants without the feature selection modules (MPM-NO-FS and MPM-NO-MPFS) or on the MPM-NO-SDNE model. Without the feature selection module, the network employed for the embeddings generation contains too many PCG association connections. As biological

data usually contains many false positives, adding all PCG associations introduces additional noise and redundancy. Similarly, without the structural embeddings (MPM-NO-SDNE), MPM only relies on the associated PCGs, miRNA, and disease semantic similarity features to generate predictions without the information about the miRNA/disease interaction patterns. The drop in performance observed in MPM's simpler variants further emphasizes the importance of our feature selection module for information filtering as well as the SDNE module for feature extraction from the raw association structural patterns.

6.4.2 MPM with different binary classifiers

TABLE 6.11: MPM with different binary classifiers results on the 18 inductive testing dataset for new diseases.

dataset	Classifier	AUC	AP	SN	SP	Acc	Pre	F1	Mcc	Top100
	SVM	**.906**	**.829**	82.2	87.9	86.7	64.2	**71.8**	64.4	97.8
	Random Forest	.895	.824	55.7	**98.6**	**89.8**	**91.4**	69.1	**66.3**	**99.4**
	AdaBoost	.892	.788	78.3	74.7	75.4	65	65.5	54	96.6
D001943	K-Nearest Neighbors	.838	.615	68.2	88.8	84.6	61.4	64.5	55	84.6
	Gaussian Naive Bayes	.511	.207	76.6	24.8	35.4	20.8	32.7	1.3	20.7
	MLP	.806	.698	69.2	68	68.3	52.5	52	39	92.5
	Decision Tree	.839	.654	**91.6**	28.1	41.1	3.6	42.6	17.5	92.9
	SVM	**.913**	.798	**90.7**	66.6	70.7	37.1	52.3	44.4	92.1
	Random Forest	.905	**.806**	74.2	95.1	**91.5**	75.6	**74.9**	**69.8**	**95.2**
	AdaBoost	.892	.77	52	95.8	88.4	82.7	55.9	55.6	93.9
D015179	K-Nearest Neighbors	.841	.514	83.9	71.8	73.8	41	54.2	45.1	65.6
	Gaussian Naive Bayes	.534	.181	86.9	19.6	31.1	18.2	30.1	6.4	18.6
	MLP	.836	.685	87.2	45.6	52.7	28.9	41.8	26.7	86.7
	Decision Tree	.864	.69	55.3	**98.0**	90.7	**86.3**	65.8	63.7	89.6
	SVM	**.949**	**.841**	70.2	97	**92.7**	81.9	**75.6**	**71.7**	95.4
	Random Forest	.938	.837	82	92.7	91	68.3	74.5	69.5	**96.0**
	AdaBoost	.926	.79	79.4	89.7	88	67.5	70.1	65.6	94.3
D013274	K-Nearest Neighbors	.808	.431	85.7	58.7	63	29.7	43.7	33.4	56.2
	Gaussian Naive Bayes	.564	.181	**90.4**	22.3	33.2	18.2	30.3	11.6	18.9
	MLP	.905	.786	40.5	**99.2**	89.8	**91.1**	55.5	56.4	92.3
	Decision Tree	.825	.562	83.4	69.2	71.5	44.6	55.2	44.7	64.9
	SVM	**.936**	**.768**	95.4	63.4	67.8	30.8	46.2	41.8	87.5
	Random Forest	.926	.764	92.3	76.5	78.7	39.4	55	51	**87.6**
	AdaBoost	.921	.731	**97.5**	20.7	31.3	20.7	32.2	13.7	85.6
D008175	K-Nearest Neighbors	.848	.585	71.2	**94.7**	**91.5**	**68.3**	**69.7**	**64.7**	78.2
	Gaussian Naive Bayes	.563	.155	91.9	20.5	3.4	15.6	26.7	10.9	17.1
	MLP	.859	.653	89.6	42.8	49.2	26.6	38.1	25.4	80.1
	Decision Tree	.815	.518	92.7	42.6	49.5	26.6	38.6	25	67.7
	SVM	**.933**	**.776**	59.1	**97.2**	**92.0**	**76.9**	66.8	63.1	**91.1**
	Random Forest	.922	.733	80.1	91.5	90	59.6	**68.3**	**63.5**	85.5
	AdaBoost	.908	.717	56.9	96.6	91.2	66.5	60	56.8	84.7
D011471	K-Nearest Neighbors	.882	.562	75.9	91	88.9	56.9	65	59.5	73.2
	Gaussian Naive Bayes	.555	.15	**87.6**	23	31.7	15.1	25.8	8.8	14.5

	MLP	.882	.669	49.1	96.8	90.3	72.9	56.4	53.8	79.6
	Decision Tree	.864	.605	78.5	83.5	82.8	58.4	64.2	57.5	77.5
	SVM	**.943**	0.800	55.4	98.4	**92.7**	83.7	66.7	64.4	90.9
	Random Forest	.934	**.802**	40	**99.5**	91.6	**92.4**	55.7	57.5	**92.1**
	AdaBoost	.938	.756	69.5	96.2	92.6	74.1	**71.1**	**67.3**	86.8
D002289	K-Nearest Neighbors	.769	.4	42.2	94.8	87.8	55.8	47.9	41.8	61.5
	Gaussian Naive Bayes	.567	.15	**92.7**	20.5	30	15.1	26	11.4	14.9
	MLP	.869	.663	62.6	90.2	86.5	62.6	58.5	53.8	80.0
	Decision Tree	.858	.604	41.1	98	90.5	81.3	51	51.6	72.6
	SVM	**.964**	**.798**	90.3	91.3	91.2	58.2	70.8	68.1	**88.9**
	Random Forest	.958	.792	87.9	92.5	**91.9**	61.1	**72.0**	**69.0**	87.1
	AdaBoost	.953	.751	84	**93**	**91.9**	**63.2**	71.6	68.4	84.6
D010051	K-Nearest Neighbors	.886	.506	88.4	76.7	78.0	34.2	49.1	45.7	60.7
	Gaussian Naive Bayes	.588	.14	**95.3**	22	30.7	14.1	24.5	14	14.7
	MLP	.932	.749	84.6	86.8	86.6	52.5	62.7	59.4	85.2
	Decision Tree	.907	.625	74.4	84.4	83.2	60.5	59.1	55.3	72.2
	SVM	**.934**	**.752**	50.5	98.2	92.6	78.7	61.4	59.3	**83.7**
	Random Forest	.917	.724	52.1	98.0	**92.6**	77.6	62.1	**59.8**	82.5
	AdaBoost	.922	.688	39.1	97.9	91.0	**84.9**	47.5	50.6	80.9
D008545	K-Nearest Neighbors	.845	.513	65.7	94.2	90.9	60.1	**62.8**	57.7	70.8
	Gaussian Naive Bayes	.564	.132	**90.6**	22.1	30.1	13.3	23.2	10.1	11.3
	MLP	.886	.641	21.9	**99.2**	90.2	80.8	32.1	36.9	75.5
	Decision Tree	.883	.577	62	95.8	91.8	70.4	62.1	59.9	71.9
	SVM	**.96**	**.781**	79.6	96.4	**94.4**	74.5	**77.0**	**73.9**	**83.8**
	Random Forest	.947	.759	49	**98.4**	92.6	80.9	59.3	58.6	81.6
	AdaBoost	.945	.741	49.4	98.2	92.5	**81.2**	58.2	58.2	82.0
D005910	K-Nearest Neighbors	.892	.622	74.5	95.7	93.2	69.8	71.9	68.2	78.0
	Gaussian Naive Bayes	.589	.139	**95.9**	21.9	30.5	14	24.4	14.3	13.5
	MLP	.892	.646	60.7	94.5	90.5	62.4	59.7	55.6	75.1
	Decision Tree	.917	.607	63.2	93.7	90.1	46.1	52.6	49.7	61.5
	SVM	**.945**	.734	70.9	96	93.3	68.7	69.8	66	79.2
	Random Forest	.938	**.736**	75.1	95.6	**93.3**	67.6	**71.1**	**67.5**	**79.7**
	AdaBoost	.935	.678	73.3	94.6	92.3	64	66.7	63.6	77.0
D005909	K-Nearest Neighbors	.879	.542	69.4	94.4	91.7	60.8	64.7	60.3	70.2
	Gaussian Naive Bayes	.583	.129	**94.9**	21.8	29.8	13	22.8	13	13.7
	MLP	.873	.537	39.8	**96.6**	90.3	61.2	45.7	43.3	62.7
	Decision Tree	.902	.558	77.1	91.3	89.7	47.5	58.5	55.9	64.0
	SVM	**.947**	.759	85.9	87.4	87.3	44.2	58.3	55.8	85.0
	Random Forest	.94	**.785**	88.3	84.7	85.1	40.1	55.1	53	**88.7**
	AdaBoost	.929	.712	89.9	60.3	63.3	33.6	45.5	38.1	84.0
D001749	K-Nearest Neighbors	.909	0.528	84.4	**91.1**	**90.4**	**52.3**	**64.6**	**61.7**	60.8
	Gaussian Naive Bayes	.56	.116	88.7	22.6	29.4	11.7	20.6	8.4	11.5
	MLP	.905	.647	73.8	90.7	89	52.2	59.7	55.9	73.9
	Decision Tree	.778	.411	**90.8**	55.6	59.3	30.8	41.3	35.8	45.4
	SVM	**.943**	**.714**	43.1	98.9	93.5	81.7	56.4	56.5	77.6
	Random Forest	.932	.713	44	**99**	**93.6**	**82.6**	57.3	57.4	**78.8**
	AdaBoost	.929	.653	48.9	98	93.2	78.1	57.5	57.3	76.1
D012516	K-Nearest Neighbors	.844	.502	52.6	97.1	92.7	66.5	58.7	55.3	68.4
	Gaussian Naive Bayes	.578	.115	**90.7**	24.6	31.1	11.6	20.6	1.8	11.1
	MLP	.862	.542	49.8	95.9	91.3	58.1	52.8	48.8	63.9

	Decision Tree	.902	.597	52.4	97.8	93.3	74.4	**60.3**	**58.5**	72.0
D010190	SVM	**.949**	**.791**	66.9	97.4	**94.4**	73.9	70.2	67.3	**84.2**
	Random Forest	.944	.749	83.1	94.2	93.1	60.9	**70.2**	**67.5**	79.7
	AdaBoost	.939	.733	47.7	98.3	93.4	**85.0**	53.7	56.3	80.7
	K-Nearest Neighbors	.921	.561	87.7	91.2	90.9	52	65.3	63.2	64.7
	Gaussian Naive Bayes	.579	.114	**94.3**	21.4	28.5	11.5	20.5	11.7	11.4
	MLP	.919	.672	39.1	**98.5**	92.7	76.5	49.8	50.4	72.2
	Decision Tree	.918	.598	86.5	91.5	91	57.5	68.3	65.9	68.9
D006333	SVM	**.956**	**.675**	70.4	95.5	93.3	60.1	**64.7**	**61.3**	**70.5**
	Random Forest	.949	.669	67.2	95.9	**93.4**	61.2	64	60.5	68.8
	AdaBoost	.951	.621	47	**97.3**	92.9	**67.2**	51.9	51.1	66.6
	K-Nearest Neighbors	.848	.49	65.7	95.3	92.7	57.3	61.2	57.3	65.9
	Gaussian Naive Bayes	.617	.112	**98.6**	24.4	30.9	11.1	20.0	15.6	11.0
	MLP	.816	0.502	46.7	96.4	92.0	57.7	50.3	47.1	60.8
	Decision Tree	.909	.585	65.8	95.8	93.1	61.4	62.4	59.4	67.9
D002292	SVM	**.944**	.662	85.3	88.9	88.6	42.5	56.7	55.1	67.6
	Random Forest	.939	**.684**	80.0	93.1	91.9	52.4	63.3	60.7	**70.8**
	AdaBoost	.921	.621	88.5	75.5	76.7	36.9	49.6	46.9	67.1
	K-Nearest Neighbors	.891	.416	86.4	80.1	80.6	29.8	44.1	43.2	51.6
	Gaussian Naive Bayes	.577	.102	**93.6**	21.5	27.8	10.2	18.4	10.6	10.4
	MLP	.913	.573	79.1	86	85.4	41.6	52.4	50.1	62.2
	Decision Tree	.897	.488	81.4	**93.1**	**92.1**	**53.1**	**64.2**	**61.8**	58.1
D003110	SVM	.943	.549	**99.6**	26.7	32.4	11	19.6	16.5	59.4
	Random Forest	**.945**	**.659**	99.2	19.6	25.8	9.7	17.7	12.6	**68.3**
	AdaBoost	.939	.636	96.2	47.5	51.3	20.2	31.8	26.9	66.5
	K-Nearest Neighbors	.597	.097	98.3	6.7	13.8	8.2	15.1	5.3	11.1
	Gaussian Naive Bayes	.578	.091	95.4	19.9	25.7	9.1	16.7	10.5	8.1
	MLP	.888	.485	98.9	10.8	17.7	8.6	15.9	8.1	53.2
	Decision Tree	.814	.361	90.7	**62.7**	**64.9**	**28.6**	**40.2**	**37.7**	39.5
D015470	SVM	.953	.64	73.8	94.3	92.9	49.1	58.9	56.6	**64.2**
	Random Forest	**.953**	**.655**	70.9	**95.3**	**93.6**	**52.8**	**60.5**	**57.8**	63.4
	AdaBoost	.944	.597	83.3	91.3	90.7	42.9	55.9	55.4	60.1
	K-Nearest Neighbors	.915	.457	80.2	92.7	91.8	44.9	57.6	56.2	54.9
	Gaussian Naive Bayes	.589	.083	**96.4**	21.2	26.4	8.3	15.4	11.2	7.6
	MLP	.911	.562	69.6	92.1	90.6	45.1	53.4	51	57.3
	Decision Tree	.921	.425	81.7	92.2	91.4	44.5	56.9	56.1	48.7
D002294	SVM	**.962**	.647	93.5	91.1	91.3	42.1	58	59.3	61.8
	Random Forest	.962	**.669**	90.6	**92.5**	**92.4**	**45.4**	**60.5**	**60.9**	**62.8**
	AdaBoost	.952	.601	86.7	91.2	90.9	43.6	56.0	56.8	60.8
	K-Nearest Neighbors	.915	.401	85.5	89.7	89.5	36.4	51.1	51.6	47.6
	Gaussian Naive Bayes	.59	.077	**97.1**	20.8	25.7	7.8	14.4	11.1	6.3
	MLP	.928	.564	83.6	89.4	89	39.5	52.3	52.4	56.5
	Decision Tree	.936	.465	91.3	90.4	90.4	40	55.4	56.6	52.1

This section presents an ablation study regarding MPM's performance with different binary classification models. In addition to the originally proposed model (with Random Forest), we also report the performance of MPM with the following classifiers: SVM [202], AdaBoost [2], K-Nearest Neighbors [101], Gaussian Naive Bayes [67], Multi-layer Perceptron [154] (MLP), and Decision Tree [43]. For all added binary classifiers, we use the default parameter sets.

Table 6.11 presents the results corresponding to MPM with different binary classifiers on the 18 test sets for new diseases, averaged over 20 experimental runs. Looking at the results, we see that the Random Forest model results in the highest Precision, F1, and Top100 scores in 11 out of the 18 test sets. These results support our choice of architecture.

6.5 Case studies

Let $\mathbf{H} = \text{HMDD2} \cup \text{HMDD3}$ denote the set of all known associations retrieved from the HMDD databases. We here present three case studies to showcase the application of MPM in realistic scenarios.

6.5.1 MPM for a disease with scarce knowledge

Down syndrome or Trisomy 21 is a condition in which a child is born with an extra copy of their 21st chromosome [178]. *Down Syndrome*'s patients usually suffer from mild-to-moderate learning disabilities[178]. According to the data deposited in the HMDD 2.0 and HMDD 3.0 databases and two recent works [60, 185], there are only 10 miRNAs known to be associated with the disease of our interest. We assume that *Down Syndrome* is a completely new disease and take similar steps as those presented in the 6.2.3 section to construct the training and testing data. In short, our training data consists of all known associations in **H** for all diseases other than the *Down Syndrome*. We test MPM on the complete test set consisting of all possible combinations between the *Down Syndrome* and 1,618 miRNAs.

TABLE 6.12: MPM's prediction scores for *Down Syndrome* and all 1,618 miRNAs.

Rank	miRNA	pred.	Rank	miRNA	pred.
	...		82	hsa-mir-125b-2	0.579253110400618
2	hsa-mir-155	0.963881105523116		...	
3	hsa-mir-146a	0.934014942433006	105	hsa-mir-99a	0.482246263067031
4	hsa-mir-16-1	0.895608127697913		...	
	...		140	hsa-mir-1246	0.404202397336283
33	hsa-mir-27b	0.689694528927961		...	
	...		261	hsa-let-7c	0.244887327696169
38	hsa-mir-27a	0.671913693062923		...	
	...		1576	hsa-mir-802	0.130087980984639

How effective is MPM in restricting and prioritizing the search space for the potentially associated miRNAs? Table 6.12 presents the average predictions made by MPM after 20 experimental runs. Though we perform the search on a complete test set of 1,618 testing samples and the training data does not contain known associations for *Down Syndrome*, 3 known-to-associate miRNAs (marked as blue in Table 6.12) already appear in the top 4 highest predicted results. The other associated miRNAs appear at 33th, 38th, 82th, 105th, 140th, 261th, and 1576th positions in the prediction list. With 3

appearing in the top 4 and 5 out of 10 known associations appearing in the top 38 of the generated prediction results, our method would significantly help restrict and prioritize the search space for wet-lab experiments.

How effective is MPM with some added domain knowledge? Since *Down Syndrome* relates to a redundant chromosome 21 copy, we retrieve the miRNA location information from miRTarBase [86] and present MPM's predicted results for all miRNAs located on chromosome 21 in Table 6.13. Blue is used to mark the associated miRNAs. Note that the model training data does not contain the association data for *Down Syndrome*.

TABLE 6.13: MPM's prediction results for *Down Syndrome* and the miRNAs that are located on chromosome 21.

rank	miRNA	pred.	rank	miRNA	pred.
1	hsa-mir-155	0.963881105523116	11	hsa-mir-4760	0.172962854437391
2	hsa-mir-125b-2	0.579253110400618	12	hsa-mir-5692b	0.168364046134056
3	hsa-mir-99a	0.482246263067031	13	hsa-mir-6508	0.163143029370321
4	hsa-let-7c	0.244887327696169	14	hsa-mir-6070	0.16232917173827
5	hsa-mir-548x	0.239129159103197	15	hsa-mir-6815	0.159395572782035
6	hsa-mir-3648-1	0.206785057828119	16	hsa-mir-8069-1	0.155993241075239
7	hsa-mir-4759	0.200771150543586	17	hsa-mir-6724-1	0.153456269809843
8	hsa-mir-3197	0.19795748172893	18	hsa-mir-6501	0.152740622433185
9	hsa-mir-6130	0.194382789321313	19	hsa-mir-6814	0.145666592873055
10	hsa-mir-4327	0.176297567535453	20	hsa-mir-802	0.130087980984639

By restricting the miRNA search space, we have much more promising prediction results, with 4 out of 5 associated miRNAs appearing at the top of the list. Adding more related domain information like chromosomal location, tissue expression profiles, etc., thus helps in restricting the miRNA search space to obtain more meaningful prediction results. Nonetheless, we release predicted association probabilities for all 1,618 miRNAs to encourage field experts' assessments as well as to enable them to perform customized subset selection without the need to retrain/rerun the model.

6.5.2 MPM for a disease with many false positives

Parkinson disease (PD) is the second most common neurodegenerative disease worldwide [115]. Existing human association studies for the *Parkinson* disease resulted in inconsistent findings with many "false positives" as reported in [191]. In this case study, we take a closer look at the generated predictions from MPM for the *Parkinson* disease. We train MPM with all the available data in **H**. More specifically, besides the data for other diseases, the training data contains 61 known associations for *Parkinson*. Among those, there are 8 true positives (those that are confirmed as positives in [191]) and 26 false positives [191] (those that are marked as positive in **H** but are confirmed as negative in [191]).

TABLE 6.14: The predicted association probabilities for the *true positive* (marked as blue) and *true negative* miRNAs [191] corresponding to the *Parkinson* disease.

rank	miRNA	pred.	rank	miRNA	pred.	rank	miRNA	pred.
1	hsa-mir-7-1	0.99	38	hsa-mir-425	0.93	75	hsa-mir-345	0.81
2	hsa-mir-30d	0.99	39	hsa-mir-10b	0.93	76	hsa-mir-142	0.8
3	hsa-mir-19b-1	0.99	40	hsa-mir-29a	0.93	77	hsa-mir-708	0.8
4	hsa-mir-146a	0.99	41	hsa-mir-99b	0.93	78	hsa-mir-1249	0.78
5	hsa-mir-335	0.99	42	hsa-mir-543	0.93	79	hsa-mir-190a	0.78
6	hsa-mir-193a	0.99	43	hsa-mir-34b	0.93	80	hsa-mir-129-1	0.77
7	hsa-mir-214	0.98	44	hsa-mir-431	0.92	81	hsa-mir-331	0.76
8	hsa-mir-141	0.98	45	hsa-mir-99a	0.92	82	hsa-mir-181c	0.75
9	hsa-mir-151a	0.98	46	hsa-mir-19a	0.92	83	hsa-mir-150	0.73
10	hsa-mir-126	0.98	47	hsa-mir-29c	0.92	84	hsa-mir-489	0.72
11	hsa-mir-7-2	0.98	48	hsa-mir-1301	0.91	85	hsa-mir-505	0.68
12	hsa-mir-146b	0.98	49	hsa-mir-30b	0.91	86	hsa-mir-203a	0.67
13	hsa-mir-29b-2	0.98	50	hsa-mir-152	0.9	87	hsa-mir-454	0.65
14	hsa-mir-30a	0.98	51	hsa-mir-125b-2	0.9	88	hsa-mir-130a	0.64
15	hsa-mir-199b	0.98	52	hsa-mir-125a	0.9	89	hsa-mir-149	0.62
16	hsa-mir-34c	0.98	53	hsa-mir-137	0.9	90	hsa-mir-1264	0.62
17	hsa-mir-132	0.98	54	hsa-mir-204	0.89	91	hsa-mir-744	0.61
18	hsa-mir-451a	0.97	55	hsa-mir-224	0.89	92	hsa-mir-301b	0.6
19	hsa-mir-133b	0.97	56	hsa-mir-148b	0.89	93	hsa-mir-154	0.59
20	hsa-mir-10a	0.97	57	hsa-mir-409	0.89	94	hsa-mir-184	0.55
21	hsa-mir-16-1	0.97	58	hsa-mir-504	0.89	95	hsa-mir-223	0.54
22	hsa-mir-30c-2	0.97	59	hsa-mir-186	0.89	96	hsa-mir-532	0.49
23	hsa-mir-127	0.96	60	hsa-mir-448	0.88	97	hsa-mir-1296	0.48
24	hsa-mir-145	0.96	61	hsa-mir-769	0.87	98	hsa-mir-873	0.44
25	hsa-mir-195	0.96	62	hsa-mir-1248	0.87	99	hsa-mir-125b-1	0.42
26	hsa-mir-497	0.96	63	hsa-mir-92a-2	0.87	100	hsa-mir-1298	0.35
27	hsa-mir-338	0.96	64	hsa-mir-328	0.86	101	hsa-mir-939	0.34
28	hsa-mir-222	0.96	65	hsa-mir-92a-1	0.86	102	hsa-mir-488	0.29
29	hsa-mir-221	0.96	66	hsa-mir-20a	0.85	103	hsa-mir-330	0.24
30	hsa-mir-22	0.96	67	hsa-mir-25	0.85	104	hsa-mir-192	0.2
31	hsa-mir-299	0.96	68	hsa-mir-23a	0.85	105	hsa-mir-626	0.19
32	hsa-mir-424	0.95	69	hsa-mir-191	0.85	106	hsa-mir-26b	0.16
33	hsa-mir-21	0.95	70	hsa-mir-140	0.84	107	hsa-mir-577	0.16
34	hsa-mir-17	0.95	71	hsa-mir-136	0.83	108	hsa-mir-654	0.15
35	hsa-mir-148a	0.94	72	hsa-mir-16-2	0.82	109	hsa-mir-378a	0.15
36	hsa-mir-143	0.94	73	hsa-mir-98	0.82	110	hsa-mir-501	0.12
37	hsa-mir-28	0.94	74	hsa-mir-27b	0.81			

We present the predicted association probabilities for all 12 *true positive* and 98 *true negative* miRNAs retrieved from the meta analysis [191] corresponding to the *Parkinson* disease in Table 6.14. Though the training data contains more than three folds of the false-positive associations (26 false positives vs. 8 true positives), we observe that all 12 true positives reported in [191] could be found in the top 50 predictions. Among those, 5 out of 12 appear in the top 8, while 8 out of 12 show up in the top 19 predictions.

These results support that MPM acquires good performance in differentiating between the true positive and true negative miRNAs even with the noisy training data.

6.5.3 Survival analysis for Precursor B-cell lymphoblastic leukemia

Precursor B-cell lymphoblastic leukemia (PBLL) is the most common type of Acute lymphoblastic leukemia that is characterized by a high number of B-cell lymphoblasts found in blood and bone marrow. According to the data deposited in the HMDD databases, there are 7 miRNAs known to be associated with PBLL. In this case study, we perform survival analysis on PBLL patients' data.

MiRNA expression and survival outcome. We download the miRNA expression and survival information for PBLL patients from TCGA Genomic Data Commons (GDC) [68] using the GDC Data Transfer Tool [69]. As a preprocessing step, we remove the patients without survival information and retain only the records that have the *Sample Type* as *Primary Tumor*. For the patients that have only one sample, the miRNA expression values are taken as the read per million values. For each patient with more than one sample, each miRNA expression value is calculated as the average of all the available reads per million values. The final preprocessed data contains the miRNA expression profiles and survival outcomes for 167 PBLL patients. For each miRNA, we use StepMiner [182] to compute a threshold that can robustly differentiate between the high and low expression levels. The computed thresholds are used to discretize the data so that the miRNA continuous expression values can be divided into high, intermediate, and low expression classes. We use the log-rank test [80, 144, 167] to assess the statistical significance of the survival difference between the high and low expression classes. The Kaplan-Meier analysis and log-rank test are performed using the *lifelines* [42] package.

MPM prediction. We train MPM with all known associations deposited in the HMDD databases for all diseases other than PBLL and generate MPM's prediction scores for all 1,618 miRNAs.

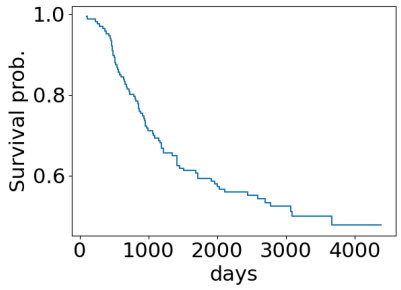


FIGURE 6.5: The Kaplan survival curve of PBLL patients.

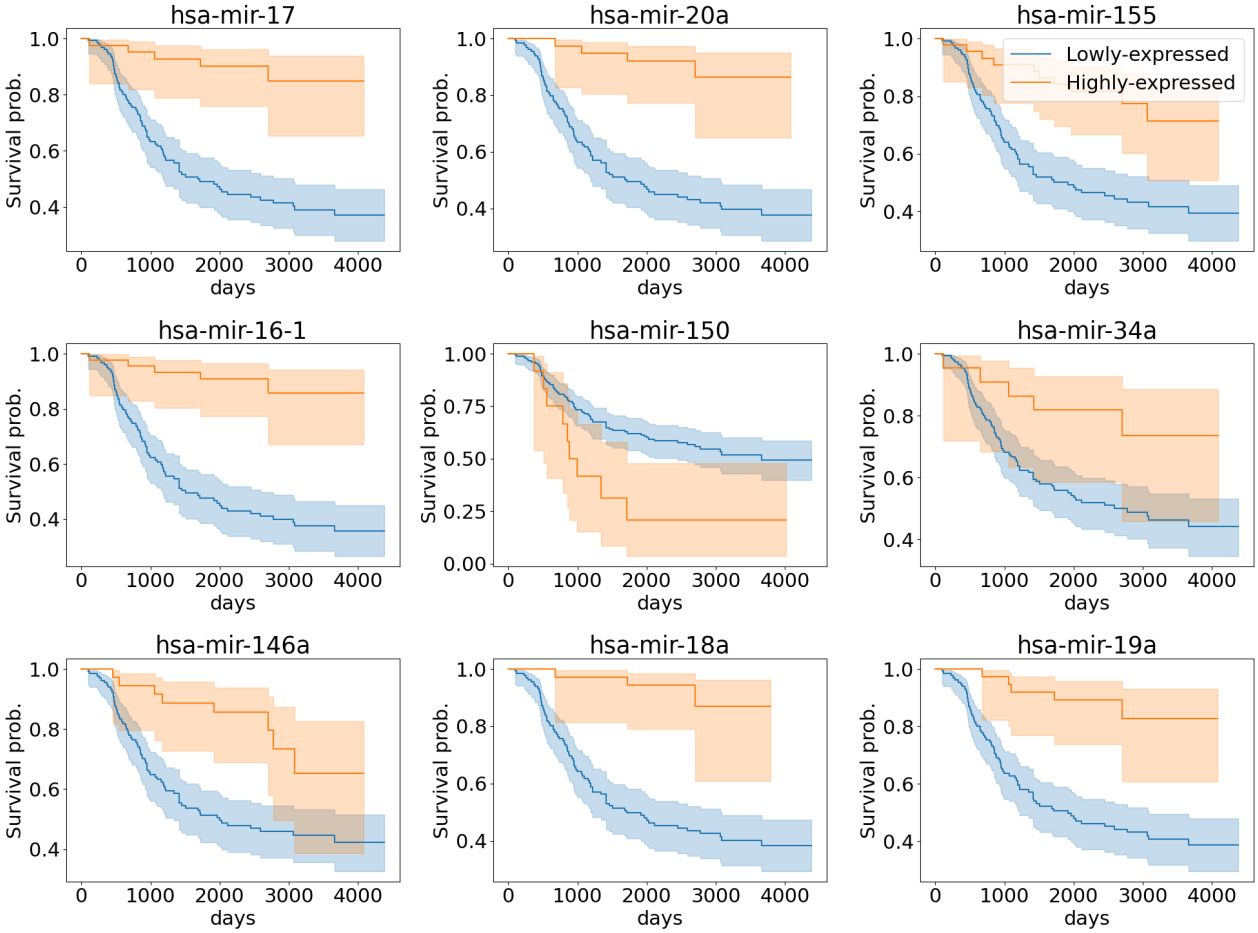


FIGURE 6.6: Kaplan–Meyer survival curves of PBLL patients stratified by the top miRNAs with the highest prediction scores.

Results. The Kaplan-Meier survival curve for PBLL patients is presented in Figure 6.5. According to the log-rank test results, we identify 310 miRNAs associated with patients' survival outcomes with a p-value < 0.05. We refer to this set as $\mathcal{L}$. We observe that none of the known-to-be-associated miRNAs (deposited in the HMDD databases) appear in $\mathcal{L}$. But from the top 10 miRNAs that have the highest prediction scores generated by MPM, 8 already appear in $\mathcal{L}$. Among the top 20 miRNAs that have the highest prediction scores, 13 already appear in $\mathcal{L}$. Table 6.15 presents the top miRNAs that have the highest prediction scores that appear in $\mathcal{L}$, along with their rank in MPM's prediction list. The full list of $\mathcal{L}$ and all MPM's prediction scores can be downloaded from `https://git.l3s.uni-hannover.de/dong/mpm/-/tree/master/PBLL`. Figure 6.6 shows the Kaplan-Meir survival curves of PBLL patients stratified by the top miRNAs that have the highest prediction scores generated by MPM. All things considered, for PBLL, MPM prediction results agree well with the survival analysis results. This further supports the applicability of MPM in identifying potential prognostic miRNAs for complex diseases.

TABLE 6.15: The top miRNAs with the highest prediction scores that appear in $\mathcal{L}$ - the list of associated miRNAs output from the survival analysis.

rank	miRNA	pred.	rank	miRNA	pred.	rank	miRNA	pred.
2	hsa-mir-17	0.98	24	hsa-mir-181a-2	0.9	51	hsa-mir-132	0.79
3	hsa-mir-20a	0.98	25	hsa-mir-19b-1	0.9	54	hsa-mir-106a	0.78
4	hsa-mir-155	0.98	27	hsa-mir-22	0.89	56	hsa-mir-378a	0.76
5	hsa-mir-16-1	0.97	29	hsa-mir-92a-1	0.86	58	hsa-mir-200c	0.75
6	hsa-mir-150	0.97	31	hsa-mir-106b	0.85	61	hsa-mir-149	0.75
7	hsa-mir-34a	0.96	33	hsa-mir-181b-1	0.85	62	hsa-mir-100	0.74
9	hsa-mir-146a	0.95	37	hsa-mir-130a	0.84	63	hsa-mir-200b	0.74
10	hsa-mir-18a	0.95	38	hsa-mir-125a	0.83	64	hsa-mir-192	0.74
14	hsa-mir-19a	0.94	40	hsa-mir-204	0.83	71	hsa-mir-16-2	0.73
15	hsa-mir-15a	0.94	45	hsa-mir-122	0.81	72	hsa-mir-98	0.73
17	hsa-mir-145	0.93	46	hsa-mir-25	0.81	73	hsa-mir-107	0.72
18	hsa-mir-143	0.92	47	hsa-mir-15b	0.81	75	hsa-mir-335	0.72
19	hsa-mir-26a-1	0.92	48	hsa-mir-148a	0.8	76	hsa-mir-26b	0.72
23	hsa-mir-31	0.91						

6.6 An integrated, easy-to-use website

We provide an easy-to-use website to query the predictions generated by our proposed model on 1,618 miRNAs and 3,679 diseases at `http://software.mpm.leibniz-ai-lab.de/`. It is important to note that the model is trained from the data corresponding to only a few hundred miRNAs and a few hundred diseases. We offer a large computational prediction capability for thousands of available diseases and miRNAs through the website. To enable a comprehensive analysis by the field experts, we also integrate the biologically related features into the application. In the following, we present details regarding the related biological features and a user guide for the web application.

6.6.1 Biological related features to support biologist justification and verification

As the associated pathway information is more intuitive compared with the list of associated PCGs, we perform pathway and functional enrichment analysis on the list of interacting/associated PCGs for each miRNA/disease and encapsulate the corresponding information into our web application. We perform pathway enrichment analysis by using the API provided by Reactome [62] and functional enrichment analysis by using the goscripts package [152]. We retain only pathways and GO terms whose p-values are smaller than 0.05.

6.6.2 The user guide

Figure 6.7 shows the start screen when opening the application tab in the web app and illustrates the main steps to use it. First, the user selects the **i) Main**

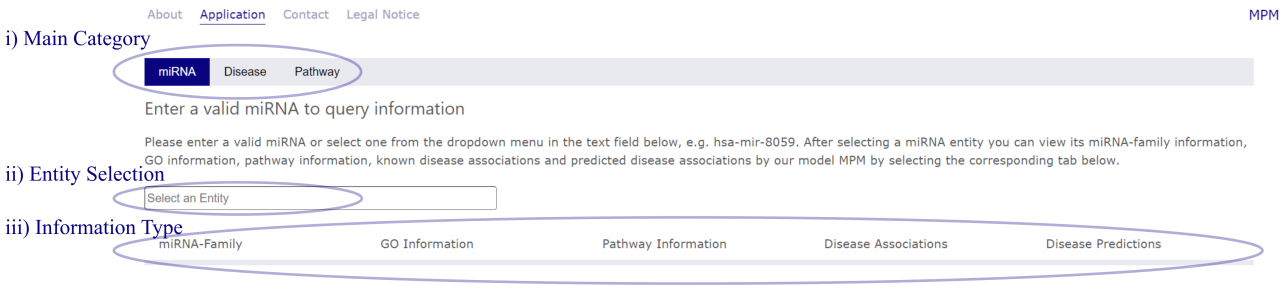


FIGURE 6.7: Web app start screen.

Category by clicking on the corresponding tab at the top of the application, i.e. *miRNA*, *Disease* or *Pathway*, marked with i) in Figure 6.7. In the next step **ii) Entity Selection**, the user selects a specific entity from that main category by either typing a valid entity name in the search field or by selecting an entity from the drop-down menu. The drop-down menu (which also serves as a search field) is marked with ii) in Figure 6.7 and opens upon selection. After a specific entity to inspect is selected, the user chooses the **iii) Information Type** they want to display by selecting the corresponding tab, marked with iii) in Figure 6.7.

Inspecting miRNAs

If the user wants to inspect a specific miRNA, they choose one of the following tabs: *miRNA-Family*, *GO information*, *Pathway Information*, *Disease Associations* (which shows confirmed associations retrieved from the HMDD databases) or *Disease Predictions* (which shows predicted association probabilities generated by MPM) tabs to query the desired information type.

- *miRNA-Family* will display all miRNAs that belong to the same family as the selected miRNA.
- *GO information* provides the GO terms (with their IDs, names, and p-values) enriched by the set of PCGs associated with the selected miRNA.
- *Pathway information* will show all pathways (with their IDs, names, and p-values) enriched by the set of PCGs associated with the selected miRNA. The GO and pathway information are sorted ascending by their p-value. The smaller the p-value, the more significant the corresponding pathway/GO term is enriched.
- The *Disease Associations* encapsulates all diseases associated with the selected miRNA that are retrieved from the HMDD databases. The MeSH ID with the corresponding disease name is provided.
- *Disease Predictions* encloses the predictions generated by MPM for the selected miRNA. Each disease record includes the disease MeSH ID, its name, and the predicted association probability, sorted in descending

order. Additionally, the column *Confirmed Association* shows if this specific association exists in the HMDD databases (indicated by '*yes*') or not (marked as '-').

Inspecting Diseases

A disease can be inspected in similar ways as those of a miRNA. Nevertheless, instead of miRNA family, the displayed information for a disease contains the *Disease Ontology*, i.e., the child and parent diseases of the selected disease entity. Figure 6.8 present the displayed screen for the *Amyloidosis* disease corresponding to the '*miRNA predictions*' tab. The predicted associated miRNAs are shown in the left column, with their corresponding confidence score in the middle column. The right column indicates whether this association was found in the HMDD databases (marked as '*yes*') or not (marked as '-').

Inspecting Pathways

When inspecting specific pathways, the user can choose between displaying the most significant miRNAs or diseases corresponding to the selected pathway entity. Figure 6.9 shows an example query for the pathway *Establishment of Sister Chromatid Cohesion*s most significant *Disease Associations*. The diseases are sorted ascending by their p-value in the right column, with the corresponding disease ID in the left and the disease name in the middle column.

6.7 Conclusion and discussion

We propose a message passing framework with multiple data sources integration, MPM, for the problem of predicting miRNA-disease associations. MPM exploits information from multiple data sources to enrich and filter the raw biologically relevant features without introducing additional parameters. Besides detecting new associations of the partially observed miRNAs or diseases, MPM can successfully generate predictions for new diseases (which has no prior observed association in the training data). Our case studies further support (i) the reliability of MPM for predicting associations for diseases with scarce knowledge and (ii) its robustness in ranking the true positives higher when many false positives are present in the training data. In addition, MPM generated predictions for the PBLL disease agree quite well with the results retrieved from survival analysis on the publicly available miRNA expression data. Besides the proposed machine learning model, we also make the generated predictions more accessible to non-expert users by encapsulating all the generated and related domain information into a publicly available website. By releasing such a user-friendly interface, we aim to foster assessments and future adoption.

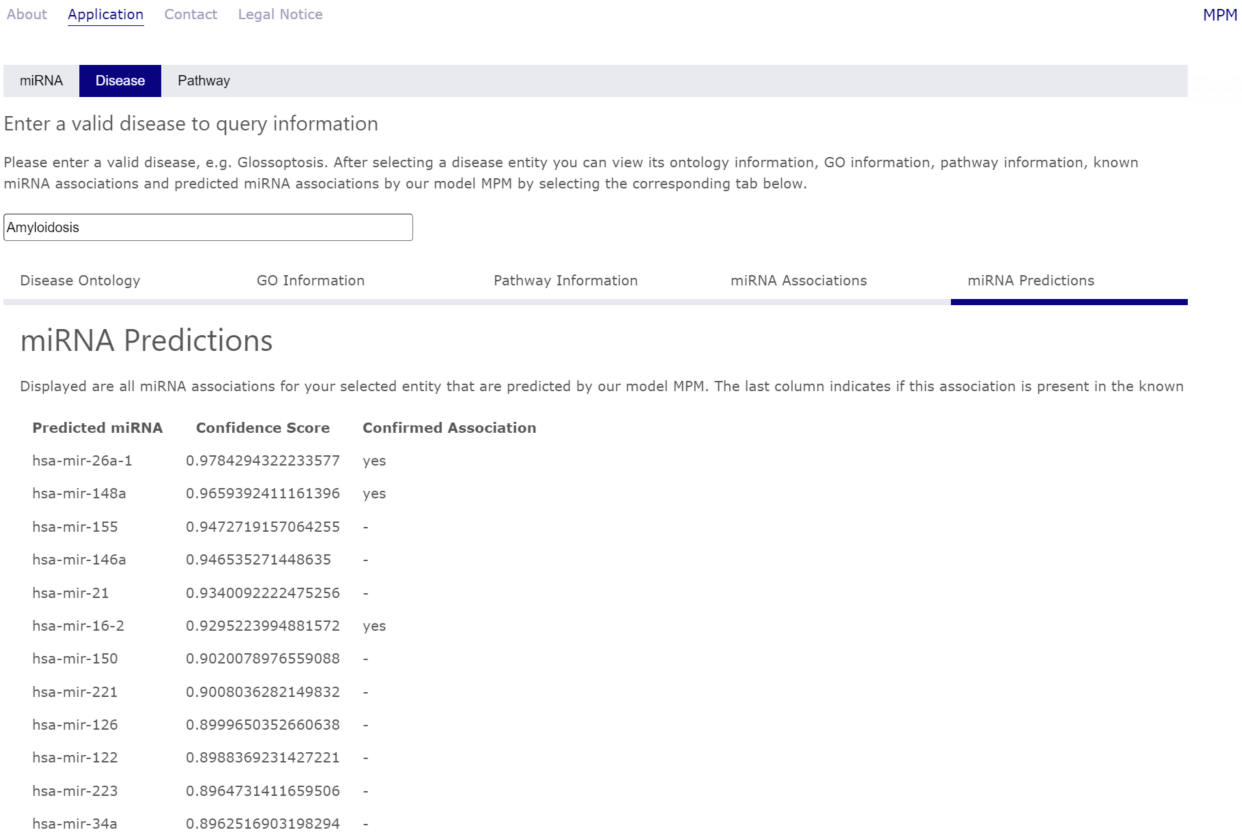


FIGURE 6.8: The Web app's encapsulated information for the *Amyloidosis* disease.

6.7.1 Potential applicability to miRNA-small molecule drug association prediction

Small molecule drugs are organic compounds with low molecular weights of around 900 Daltons. Small molecules form the majority of existing drugs and can be rapidly diffused across cell membranes [36]. Identification of miRNA-small molecule drug associations can help in disease therapy development. One of the first machine learning-based models for miRNA-small molecule drug association prediction is proposed by Jamal et al. [91]. The authors present a traditional machine learning approach that represents each miRNA-small molecule drug pair as a concatenated feature vector of miRNA and small molecule drug integrated similarities. The feature representations are then used as input to the Random Forest based binary classifier. More recent methods usually involve the use of graph representation learning techniques [78, 125, 140, 170, 171, 195, 214, 217], kernel methods [215] and matrix factorization [222]. A recent review about miRNA-small molecule drug association can be found in [36].

One shared characteristic of existing works is the utilization of small molecule drug and miRNA pre-calculated similarities. Though these works usually combine various similarities to mitigate bias and lack of information, they still suffer from issues related to the use of pre-calculated similarities, such

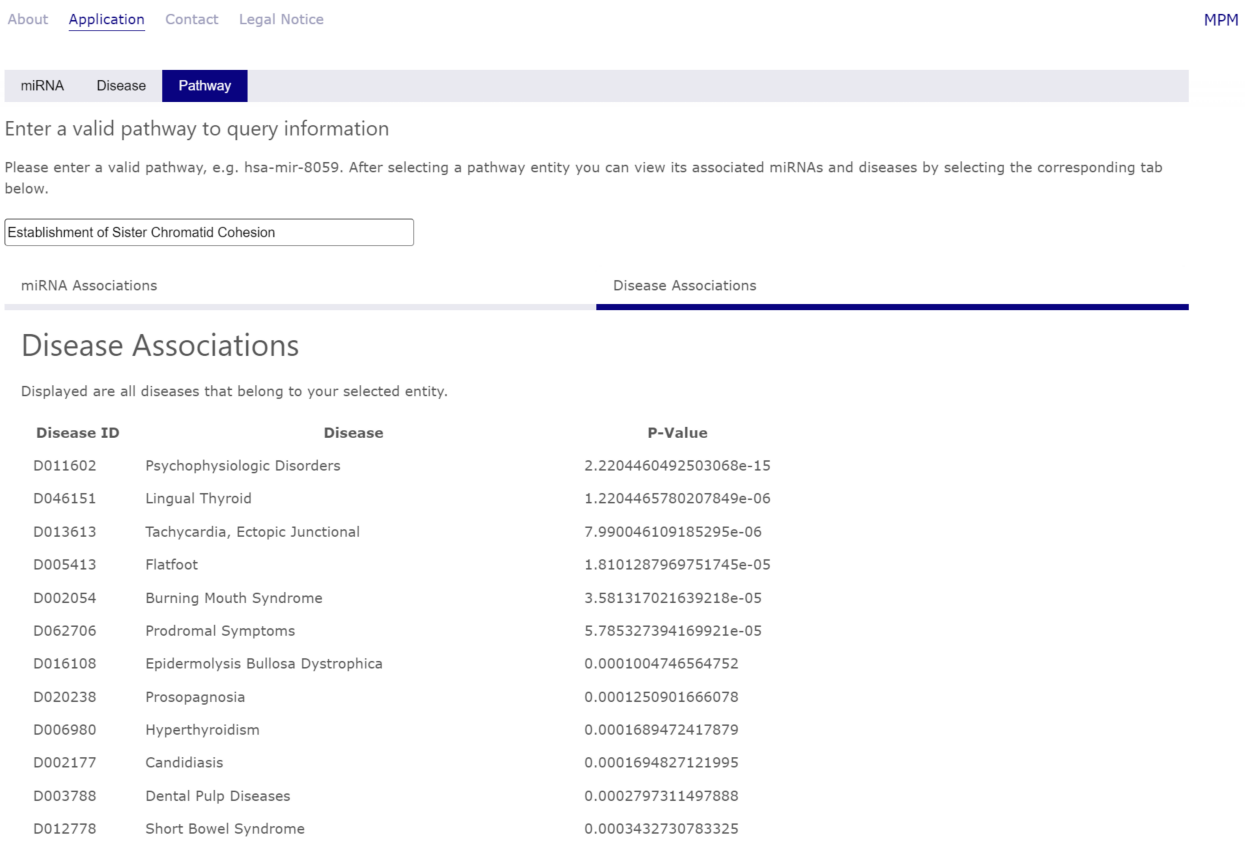


FIGURE 6.9: The Web app's encapsulated information for the *Establishment of Sister Chromatid Cohesion* pathway.

as being hard to update and maintain [52]. Graph-based methods additionally encapsulate raw miRNA-small molecule drug associations in the constructed network but the number of known associations is usually too small compared to the similarity connections. This prevents the model from learning informative association patterns. Overall, it is essential to perform task specific information filtering to remove noise and balance the amount of side information added.

Our model architecture can be easily adapted for the miRNA-small molecule drug association prediction problem. The types of input information as utilized by our model are also available for this problem. Firstly, one can extract small molecule drug similarity features based on side effects [76], functional consistency [139], chemical structure [82], and indication phenotype [76]. Secondly, we can retrieve small molecule drug-targeted genes from public databases like DrugBank [228]. Finally, each small molecule drug is also assigned to one or more ATC codes [56], which incorporate information such as its anatomical distribution, therapeutic effects, and structural characteristics. Such ATC codes are also organized into a hierarchy with different levels of granularity, like the disease ontology in our case. Nevertheless, there are still some open questions and considerations regarding (i) the choice

of similarity features, (ii) the biological rationale(s) for adding PCG associations as intermediate connecting points, and (iii) the most suitable supervised problem for performing feature selection (for example, should one use drug first level ATC code classification?). Answering such questions would require an in-depth understanding of the problem. Compared with the existing approaches, one advantage of our proposed model is that it offers a parameter-free information filtering mechanism to filter out redundant connections. High-quality input enables us to learn meaningful association patterns from the input network. Also, to the best of our knowledge, the SDNE method employed by MPM has never been used in existing works for miRNA-small molecule drug association prediction.

Chapter 7

Predicting virus-human protein-protein interaction

This chapter presents another application of our joint learning approaches on the protein-protein interaction (PPI) prediction problem. The chapter is based on our journal article: "*A multitask transfer learning framework for the prediction of virus-human protein-protein interactions*" published in BMC Bioinformatics, 2021.

7.1 Introduction

Virus infections cause an enormous and ever increasing burden on healthcare systems worldwide. The ongoing COVID-19 pandemic caused by the zoonotic virus, SARS-CoV-2, has resulted in enormous socio-economic losses [166]. Viruses infect all life forms and require host cells to complete their replication cycle by utilizing the host cell machinery. Virus infection involves several types of protein-protein interactions (PPIs) between the virus and its host. These interactions include the initial attachment of virus coat or envelope proteins to host membrane receptors, hijacking of the host translation and intracellular transport machineries resulting in replication, assembly and subsequent release of virus particles [13, 71, 197]. Besides providing mechanistic insights into the biology of infection, knowledge of virus-host interactions can point to essential events needed for virus entry, replication, or spread, which can be potential targets for the prevention, or treatment of virus-induced diseases [181].

In vitro experiments based on yeast-two hybrid (Y2H), ligand-based capture MS, proximity labeling MS, and protein arrays have identified tens of thousands of virus-human protein interactions [8, 70, 77, 92, 120, 138, 198, 226, 238]. These interaction data are deposited in publicly available databases including InAct [106], VirusMetha [19], VirusMINT [24], and HPIDB [5], and others. However, experimental approaches to unravel PPIs are limited by several factors, including the cost and time required, the generation, cultivation and purification of appropriate virus strains, the availability of recombinantly expressed proteins, generation of knock in or overexpression cell lines, availability of antibodies and cellular model systems. Computational approaches can assist *in vitro* experimentation by providing a list of most

probable interactions, which actual biological experimentation techniques can falsify or verify.

In this chapter, we cast the problem of predicting virus-human protein interactions as a binary classification problem and focus specifically on emerging viruses that has limited experimentally verified interaction data.

7.1.1 Key Challenges in learning to predict virus-Human PPI

Limited interaction data. One of the main challenges in tackling the current task as a learning problem is the *limited training data*. Towards predicting virus-host PPI, some known interactions of other human viruses collected from wet-lab experiments are employed as training data. The number of known PPIs is usually too small and thus, not representative enough to ensure the generalizability of trained models. In effect, the trained models might overfit the training data and would give inaccurate predictions for any given new virus.

Difference to other pathogens. A natural strategy to overcome the limitation posed by scarce virus protein interaction data is to employ transfer learning from available intra-species PPI or PPI data for other types of pathogens. This may, in its simplest fashion, not be a viable strategy as virus proteins can differ substantially from human or bacterial proteins. Typically, they are highly structurally and functionally dynamic. Virus proteins often have multiple independent functions so that they cannot be easily detected by common sequence-structure comparison [73, 175, 176]. Besides, virus protein sequences of different species are highly diverse [58]. Consequently, models trained for intra-species human PPI [29, 130, 131, 188, 201] or for other pathogen-human PPI [11, 47, 79, 124, 149, 200] cannot be directly used to predict virus-human protein interactions.

Limited information on structure and function of virus proteins. While for human proteins, researchers can retrieve information from many publicly available databases to extract features related to their function, semantic annotation, domains, structure, pathway association, and intercellular localization, such information is not readily available for most virus proteins. Protein crystal structures are available for some virus proteins. However, for many, predictive structures based on the amino acid sequence must be used. Thus, for the majority of virus proteins, currently, the only reliable source of virus protein information is its amino acid sequence. *Learning effective representations* of the virus proteins, therefore, is an important step towards building prediction models. Heuristics such as K-mer amino acid composition are bound to fail as it is known that virus proteins with completely different sequences might show similar interaction patterns.

7.1.2 Our Contributions

In this work, we develop a machine learning model which overcomes the above limitations in two main steps, which are described below.

Transfer Learning via Protein Sequence Representations. Though the training data on interactions as well as the input information on protein features are limited, a large number of unannotated protein sequences are available in public databases like UniProt. Inspired by advancements in Natural Language Processing, Alley et al. [3] trained a deep learning model on more than 24 million protein sequences to extract statistically meaningful representations. These representations have been shown to advance the state-of-the-art in protein structure and function prediction tasks. Rather than using handcrafted protein sequence features, we use the pre-trained model by [3] (referred to as UNIREP) to extract protein representations. The idea here is to exploit transfer learning from several million sequences to our scant training data.

Incorporating Domain Information. We further fine-tune UNIREP's globally trained protein representations using a simple neural network whose parameters are learned using a multitask objective. In particular, besides the main task, our model is additionally regularized by another objective, namely predicting interactions among human proteins. The additional objective allows us to encode (human) protein similarities dictated by their interaction patterns. The rationale behind encoding such knowledge in the learnt representation is that the human proteins sharing similar biological properties and functions would also exhibit similar interacting patterns with viral proteins. Using a simpler model and an additional side task helps us overcome overfitting, which is usually associated with models trained with small amounts of training data.

We refer to our model as MULTITASK TRANSFER (MTT) and is further illustrated in Section **??**. To sum up, we make the following contributions.

- We propose a new model that employs a transfer learning-based approach to first obtain the statistically rich protein representations and then further refines them using a multitask objective.
- We evaluated our approach on several benchmark datasets of different types for virus-human and bacteria-human protein interaction prediction. Our experimental results (c.f. Section 7.5) show that MTT outperforms several baselines even on datasets with rich feature information.
- Experimental results on the SARS-COV-2 virus receptor shows that our model can help researchers to reduce the search space for yet unknown virus receptors effectively.
- We release our code for reproducibility and further development at `https://git.l3s.uni-hannover.de/dong/multitask-transfer`.

7.2 Related work

Existing works mainly cast the PPI prediction task as a supervised machine learning problem. Nevertheless, the information about non-interacting protein pairs is usually not available in public databases. Therefore, researchers

can only either adapt models to learn from only positive samples or employ certain negative sampling strategy to generate negative examples for training data. Since the quality and quantity of the generated negative samples would significantly affect the outcome of the learned models, the authors in [124, 158, 159] proposed models that only learned from the available known positive interactions. Nourani et al. [158] and Li et al. [124] treated the virus-human PPI problem as a matrix completion problem in which the goal was to predict the missing entries in the interaction matrix. Nouretdinov et al. [159] use a conformal method to calculate p-values/confidence level related to the hypothesis that two proteins interact based on similarity measures between proteins.

Another line of work which casts the problem as a binary classification task focussed on proposing new negative sampling techniques. For instance, Eid et al [58] proposed Denovo - a negative sampling technique based on virus sequence dissimilarity. Mei et al. [150] proposed a negative sampling technique based on one class SVM. Basit et al. [11] offered a modification to the Denovo technique by assigning sample weights to negative examples inversely proportional to their similarity to known positive examples during training.

Dick et al. [47] utilizes the interaction pattern from intra-species PPI networks to predict the inter-species PPI between human-HIV-1 virus and human. Though the results are promising, this cannot be directly applied to completely new viruses where information about closely-related species is not available or to viruses whose intra-species PPI information is not available.

The works presented in [41, 45, 46, 107, 137, 141, 246] employed different feature extraction strategies to represent a virus-human protein pair as a fixed-length vector of features extracted from their protein sequences. Instead of hard-coding sequence feature, Yang et al. [234] and Lanchantin et al. [118] proposed embedding models to learn the virus and human proteins' feature representations from their sequences. However, their training data was limited to around 500,000 protein sequences. Though not very common, other types of information/features were also used in some proposed models besides sequence-based features. Those include protein functional information (or GO annotation) as in [136], proteins domain-domain associations information as in [10], protein structure information as in [79, 119], and the disease phenotype of clinical symptoms as in [136]. One limitation of these approaches is that they cannot be generalized to novel viruses where such kind of information is not available.

Among the network-based approaches, Liu et al. and Wang et al. [134, 223] constructed heterogeneous networks to compute virus and human proteins features. Nodes of the same type were connected by either weighted edges based on their sequence similarity or a combination of sequence similarity and Gaussian Interaction Profile kernel similarity. Deng et al. [45] proposed a deep-learning-based model with a complex architecture of convolutional and LSTM layers to learn the hidden representation of virus and

human proteins from their input sequence features along with the classification problem. Despite the promising performance, those studies still have the limitation posed by hand-crafted protein features.

7.3 Method

We first provide a formal problem statement.

Problem Statement. We are given protein sequences corresponding to infectious viruses and their known interactions with human proteins. Given a completely new (novel) virus, its set of protein(s) V along with its (their) sequence(s), we are interested in predicting potential interactions between V and the human proteins.

We cast the above problem as that of binary classification. The positive samples consist of pairs of virus and human proteins whose interaction has been verified experimentally. All other pairs are considered to be non-interacting and constitute the negative samples. In Section 7.4, we add details on positive and negative samples corresponding to each dataset.

Summary of the approach. The schematic diagram of our proposed model is presented in Figure 7.1. As shown in the diagram, the input to the model is the raw human and virus protein sequences which are passed through the UniRep model to extract low dimensional vector representations of the corresponding proteins. The extracted embeddings are then passed as initialization values for the embedding layers. These representations are further fine-tuned using the Multilayer Perceptron (MLP) modules (shown in blue). The fine-tuning is performed while learning to predict an interaction between two human proteins (between proteins A and B in the figure) as well as the interaction between human and virus proteins (between proteins B and C). In the following, we describe in detail the main components of our approach.

7.3.1 Extracting protein representations

Significance of using protein sequence as input. We note that the protein sequence determines the protein's structural conformation (fold), which further determines its function and its interaction pattern with other proteins. However, the underlying mechanism of the sequence-to-structure matching process is very complex and cannot be easily specified by hand-crafted rules. Therefore, rather than using hand-crafted features extracted from amino acid sequences, we employ the pre-trained UniRep model [3] to generate latent representations or protein embeddings. The protein representations extracted from UniRep model are empirically shown to preserve fundamental properties of the proteins and are hypothesized to be statistically more robust and generalizable than hand-crafted sequence features.

UniRep for extracting sequence representations. In particular, UniRep consists of an embedding layer that serves as a lookup table for each amino

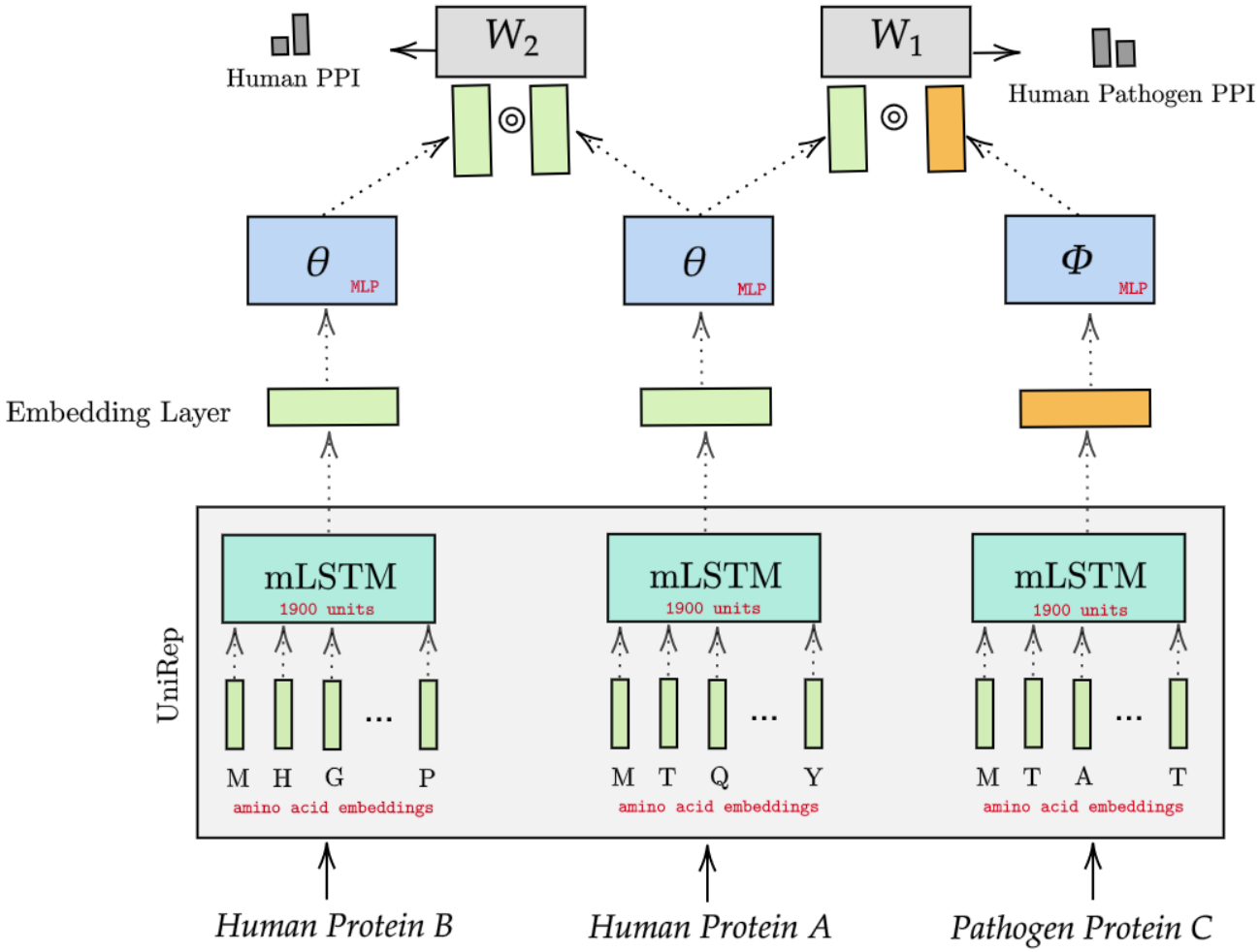


FIGURE 7.1: Our proposed MTT model for the virus-human PPI prediction problem.

acid representation. Each amino acid is represented as an embedding vector of 10 dimensions. Each input protein sequence of length N will be denoted as a two-dimensional matrix of size Nx10. That two-dimensional matrix will then feed as input to a Multiplicative Long Short Term Memory (mLSTM) network of 1900 units. The 1900 dimension is selected experimentally from a pool of architectures that require different numbers of parameters as described in [16], namely, a 1900-dimensional single layer multiplicative LSTM (∼18.2 million parameters), a 4-layer stacked mLSTM of 256 dimensions per layer (∼1.8 million parameters), and a 4-layer stacked mLSTM with 64 dimensions per layer (∼0.15 million parameters). The output from mLSTM is a 1900 dimensional embedding vector that serves as the pre-trained protein embedding for the input protein sequence. We use the calculated pre-trained virus and human protein embeddings to initialize our embedding layers. The two supervised PPI prediction tasks will further fine-tune those embeddings during training.

7.3.2 Learning framework

We further fine-tune these representations by training two simple neural networks (single layer MLP with ReLu activation) using an additional objective of predicting human PPI in addition to the main task. More precisely, the UNIREP representations will be passed through one hidden layer MLPs with ReLU activations to extract the latent representations. Let $\mathbf{X}$ denote the embedding lookup matrix. The ith row corresponds to the embedding vector of node i. The final output from MLP layers for an input v is then given by $\mathbf{hid}(v) = \text{MLP}(\mathbf{X}(v))$. To predict the likelihood of interaction between a pair (v_1, v_2) we first perform an element-wise product of the corresponding hidden vectors (output of MLPs) and pass it through a linear layer followed by sigmoid activation. In the following we provide a detailed description of our multitask objective.

Training using a multitask Objective

Let Θ, Φ denote the set of learnable parameters corresponding to fine-tuning components (as shown in Figure 7.1 in green and blue boxes), i.e., the Multilayer Perceptrons (MLP) corresponding to the virus and human proteins, respectively. Let $\mathbf{W}_1, \mathbf{W}_2$ denote the two learnable weight matrices (parameters) for the linear layers (as depicted in gray boxes in the Figure). We use VH, and HH to denote the training set of virus-human, human-human PPI, correspondingly. We use binary cross entropy loss for predicting virus-human PPI predictions, as given below:

$$\mathcal{L}_1 = \sum_{(v,h)\in VH} -z_{vh} \log y_{vh}(\Theta, \Phi, \mathbf{W}_1) - (1 - z_{vh}) \log(1 - y_{vh}(\Theta, \Phi, \mathbf{W}_1)), \tag{7.1}$$

where variables z_{vh} is the corresponding binary target variable and y_{vh} is the predicted likelihood of observing virus-human protein interaction, i.e.,

$$y_{vh}(\Theta, \Phi, \mathbf{W}_1) = \sigma((\mathbf{hid}(v) \odot \mathbf{hid}(h))\mathbf{W}_1), \tag{7.2}$$

where $\sigma(x) = 1/1 + e^{-x}$ is the sigmoid activation and $\odot$ denotes the element-wise product.

For human PPI, we predict the confidence score of observing an interaction between two human proteins. More specifically, we directly predict $z_{hh'}$ - the normalized confidence scores for interaction between two human proteins as collected from STRING [203] database. Predicting the normalized confidence scores helps us overcome the issues with defining negative interactions. We use mean square error loss to compute the loss for the human PPI prediction task as below where $y_{hh'}$ is computed similar to (7.2) for human proteins and N is the number of (h, h') pairs.

$$\mathcal{L}_2 = \frac{1}{N} \sum_{(h,h')\in HH} (y_{hh'}(\Theta, \mathbf{W}_2) - z_{hh'})^2 \tag{7.3}$$

We use a linear combination of the two loss functions to train our model.

$$\mathcal{L} = \mathcal{L}_1 + \alpha \cdot \mathcal{L}_2 \tag{7.4}$$

where α is the human PPI weight factor.

7.4 Data Description and Experimental set up

We commence by describing the 13 datasets used in this work to evaluate our approach. For all tables, $|E^+|$, $|E^-|$ refer to the number of positive and negative interactions, respectively. $|V^h|$, $|V^v|$, and $|V^b|$ denote the number of human, virus and bacteria proteins, correspondingly.

7.4.1 The realistic host cell-virus testing datasets

The NOVEL H1N1 and NOVEL EBOLA datasets. We retrieve the curated or experimentally verified PPIs between virus and human from four databases: APID [4], IntAct [106], VirusMetha [19], and UniProt [40] using the PSICQUIC web service [6]. In total, there are 11,491 known PPIs between 246 viruses and humans. From this source of data, we generate new training and testing data for the two viruses: the human H1N1 Influenza virus and Ebola virus. We name the two datasets NOVEL H1N1 and NOVEL EBOLA according to the virus present in the testing set. The positive training data for the NOVEL H1N1 dataset includes PPIs between human and all viruses except H1N1. Similarly, the positive training data for the NOVEL EBOLA dataset includes PPIs between human and all viruses except Ebola. The positive testing data for the human-H1N1 dataset contains PPIs between human and 11 H1N1 virus proteins. Likewise, the positive testing data for the human-Ebola dataset contains PPIs between human and three of the eight Ebola virus proteins (VP24, VP35, and VP40).

Negative sampling techniques such as the dissimilarity-based method [58], the exclusive co-localization method [145, 148] are usually biased as they restrict the number of tested human proteins. It is also unrealistic for a new virus because information about such restricted human protein set, generated from filtering criteria based on the positive instances, is typically unavailable. For those reasons, we argue that random negative sampling is the most appropriate, unbiased approach to generate negative training/testing samples. Since the exact ratio of positive:negative is unknown, we conducted experiments with different negative sample rates. In our new virus-human PPI experiments, we try four negative sample rates: [1,2,5,10]. In addition, to reduce the bias of negative samples, the negative sampling in the training and testing set is repeated ten times. In the end, for each dataset, we test each method with 4x4x10 = 160 different combinations of negative training and negative testing sets (with fixed positive training and test samples). The statistics for our new testing datasets are given in Table 7.1.

TABLE 7.1: The virus-human PPI realistic benchmark datasets' statistics.

	TRAINING DATA				TESTING DATA			
	$\|E^+\|$	$\|E^-\|$	$\|V^h\|$	$\|V^v\|$	$\|E^+\|$	$\|E^-\|$	$\|V^h\|$	$\|V^v\|$
NOVEL H1N1	10,858	*varies*	7,636	641	381	*varies*	622	11
NOVEL EBOLA	11,341	*varies*	7,816	659	150	*varies*	290	3
ZHOU'S H1N1	10,858	10,858	7,636	641	381	381	622	11
ZHOU'S EBOLA	11,341	11,341	7,816	659	150	150	290	3
2697049	24,698	246,980	16,638	1,066	278	448,651	16,627	27
333761	23,892	238,920	16,638	1,070	534	132,482	16,627	8
2043570	24,372	243,720	16,638	1,085	309	66,199	16,627	4
644788	24,825	248,250	16,638	1,090	54	33,200	16,627	2

The DEEPVIRAL [136] Leave-One-Species-Out (LOSO) benchmark datasets. The data was retrieved from the HPIDB database [5] to include all *Pathogen-Host* interactions that have confidence scores available and are associated with an existing virus family in the NCBI taxonomy [63]. After filtering, the dataset includes 24,678 positive interactions and 1,066 virus proteins from 14 virus families. We follow the same procedure as mentioned in [136] to generate the training and testing data corresponding to four virus species with taxon IDs: 644788 (Influenza A), 333761 (HPV 18), 2697049 (SARS-CoV-2), 2043570 (Zika virus). From now on, we will use the NCBI taxon ID of the virus species in the testing set as the dataset name. For each dataset, the positive testing data consists of all known interactions between the test virus and the human proteins. The negative testing data consists of all possible combinations of virus and 16,627 human proteins in Uniprot (with a length limit of 1000 amino acids) that do not appear in the positive testing set. Similarly, the positive training data consists of all known interactions between human protein and any virus protein, except for the one which is in the testing set. The negative training data is generated randomly with the positive:negative rate of 1:10 from the pool of all possible combinations of virus and 16,627 human proteins that do not appear in the positive training set. Statistics of the datasets are presented in table 7.1. Though performing a search on the set of 16,627 human proteins might not be a fruitful realistic strategy, we still keep the same training and testing data as released in the DEEPVIRAL study in our experiments to have a direct and fair comparison with the DEEPVIRAL method.

7.4.2 The widely used new virus-human PPI prediction benchmarked datasets.

The two datasets released by Zhou et al. [246] are widely used by recent papers to evaluate state-of-the-art models on new virus-human PPI prediction tasks. We refer to them as ZHOU'S H1N1 and ZHOU'S EBOLA where each dataset was named after the viruses in the testing sets. ZHOU'S H1N1 and ZHOU'S EBOLA share similar positive training and testing samples with the

NOVEL H1N1 and NOVEL EBOLA datasets. However, they differ in the negative training and testing samples sets. While the negative samples in NOVEL H1N1 and NOVEL EBOLA were generated randomly from the pool of all possible pairs, the negative training/testing samples in ZHOU'S H1N1 and ZHOU'S EBOLA were generated based on the protein sequence dissimilarity score. Therefore, ZHOU'S H1N1 and ZHOU'S EBOLA have the limitations as mentioned in section 7.4.1 and are not ideal for evaluating the new virus-human PPI prediction task. The data statistics for these two datasets are shown in Table 7.1.

7.4.3 The specialized testing datasets

The dataset with protein motif information (DENOVO SLIM [58]). The DENOVO SLIM dataset Virus-human PPIs were collected from VirusMentha database [19]. The presence of Short Linear Motif (SLiM) in virus sequences was used as a criterion for data filtering. SLiMs are short, recurring patterns of protein sequences that are believed to mediate protein-protein interaction [48, 156]. Therefore, sequence motifs can be a rich feature set for virus-human PPI prediction tasks. The test set [58] contained 425 positives and 425 negative PPIs (Supplementary file S12 used in DeNovo's study ST6). The training data consisted of the remaining PPI records and comprised of 1590 positive and 1515 negative records for which virus SLiM sequence is known and 3430 positives and 3219 negatives without virus SLiM sequences information. DENOVO_SLIM negative samples were also generated using the Denovo negative sampling strategy (based on sequence dissimilarity).

The BARMAN'S dataset [10] with protein domain information. The dataset was retrieved from VirusMINT database [24]. Interacting protein pairs that did not have any "InterPro" domain hit were removed. In the end, the dataset contained 1,035 positives and 1,035 negative interactions between 160 virus proteins of 65 types and 667 human proteins. 5-Fold cross-validation was then employed to test each method's performance.

7.4.4 The bacteria human PPI prediction datasets.

We evaluate our method on three datasets for three human pathogenic bacteria: BACILLUS ANTHRACIS (B1), YERSINIA PESTIS (B2), and FRANCISELLA TULARENSIS (B3), which were shared by Fatma et al. [58].

The data was first collected from HPIDB [5]. B1 belongs to a bacterial phylum different from that of B2 and B3, while B2 and B3 share the same class but differ in their taxonomic order. B1 has 3057 PPIs, B2 has 4020, and B3 has 1346 known PPIs. A sequence-dissimilarity-based negative sampling method was employed to generate negative samples. For each bacteria protein, ten negative samples were generated randomly. Each of the bacteria was then set aside for testing, while the interactions from the other two bacteria were used for training. For simplicity, we use the name of the bacteria in the

testing set as the name of the dataset. The statistics for those three datasets are presented in table 7.2.

TABLE 7.2: The bacteria-human PPI benchmark datasets' statistics.

	TRAINING DATA				TESTING DATA			
	$\|E^+\|$	$\|E^-\|$	$\|V^h\|$	$\|V^b\|$	$\|E^+\|$	$\|E^-\|$	$\|V^h\|$	$\|V^b\|$
BACILLUS ANTHRACIS	5366	15590	1559	2674	3057	9440	944	1705
YERSINIA PESTIS	4403	12880	1288	2278	4020	12150	1215	2147
FRANCISELLA TULARENSIS	7077	21590	2159	3041	1346	3440	344	1023

7.4.5 Description of Compared Methods

We compare our method with the following seven baseline methods and two simper variants of our model.

- GENERALIZED [246]: It is a generalized SVM model trained on hand-crafted features extracted from protein sequence for the novel virus-human PPI task. Each virus-human pair is represented as a vector of 1,175 dimensions extracted from the two protein sequences.
- HYBRID [45]: It is a complex deep model with convolutional and LSTM layers for extracting latent representation of virus and human proteins from their input sequence features and is trained using L1 regularized Logistic regression.
- DOC2VEC [234]: It employs the doc2vec [122] approach to generate protein embeddings from the corpus of protein sequences. A random forest model is then trained for the PPI prediction task.
- MOTIFTRANSFORMER [118]: It is a transformer-based deep neural network that pre-trains protein sequence representations using unsupervised language modeling tasks and supervised protein structure and function prediction tasks. These representations are used as input to an order-independent classifier for the PPI prediction task.
- DENOVO [58]: This model trains an SVM classifier on a hand-crafted feature set extracted from the K-mer amino acid composition information using a novel negative sampling strategy. Each protein pair is represented as a vector of 686 dimensions.
- DEEPVIRAL [136]: It is a deep learning-based method that combines information from various sources, namely, the disease phenotypes, virus taxonomic tree, protein GO annotation, and proteins sequences for intra- and inter-species PPI prediction.
- BARMAN [10]: It used an SVM model trained on a feature set consisting of the protein domain-domain association and methionine, serine, and valine amino acid composition of viral proteins.

- 2 simpler variants of MTT: Towards ablation study, we evaluate two simpler variants: (i) SINGLETASK TRANSFER (STT), which is trained on a single objective of predicting pathogen-human PPI. STT is basically the MTT without the human PPI prediction side task and (ii) NAIVE BASELINE, which is a Logistic regression model using concatenated human and pathogen protein UNIREP representations as input.

7.4.6 Implementation details and parameter set up

We use Pytorch [164] to implement our model and run it on an Nvidia GTX 1080-Ti with 11GB memory. We use Adam optimizer [109] for the model parameter optimization. For all datasets, we left out 10% of the training data for validation and performed a grid search for the best combination of parameters on that validation set. For datasets other than NOVEL H1N1 and NOVEL EBOLA, we perform parameter grid searching with the MLP hidden dimension *hid* in [8, 16,32, 64], α in $[10^{-3}, 10^{-2}, 10^{-1}, 1]$, the number of *epochs* from 0 to 200 with a step of 2 and the learning rate lr in $[10^{-3}, 10^{-2}]$. For the NOVEL H1N1 and NOVEL EBOLA datasets, we test each with 160 different combinations of negative training and negative testing. Therefore, we fix the hidden dimension to 16, $\alpha = 10^{-3}$, $lr = 10^{-3}$ and only perform grid searching on the number of epochs. The reported results for each dataset are the results corresponding to the best-performed model on the validation set.

For the DOC2VEC model, we use the released code shared by the authors with the given parameters. For the GENERALIZED and DENOVO models, we re-implement the methods in Python using all the parameters and feature set as described in the original papers. For BARMAN and DEEPVIRAL, the results are taken from the original papers or calculated from the given model prediction scores.

7.4.7 Evaluation metrics

For all benchmark datasets except the case study, we report five metrics: the Area under Receiver Operating Characteristic curve (AUC) and the area under the precision-recall curve (AP), the PRECISION, RECALL, and F1 scores.

For the case study, we report the topK score with K from 1 to 10. TopK is equal to 1 if the human receptor for SARS-COV-2 virus appears in the top K proteins that have the highest scores predicted by the model and 0 otherwise.

7.5 Result Analysis

In the following four subsections, we provide a detailed comparison of MTT with (i) methods employing hand-crafted input features, (ii) sequence embedding-based methods, (iii) an approach that uses protein domain information, (iv) simpler variants of MTT as ablation studies respectively. All statistical test results present in this section are those from the pair-wise t-test [225] on the F1 scores attained from multiple runs on the same dataset.

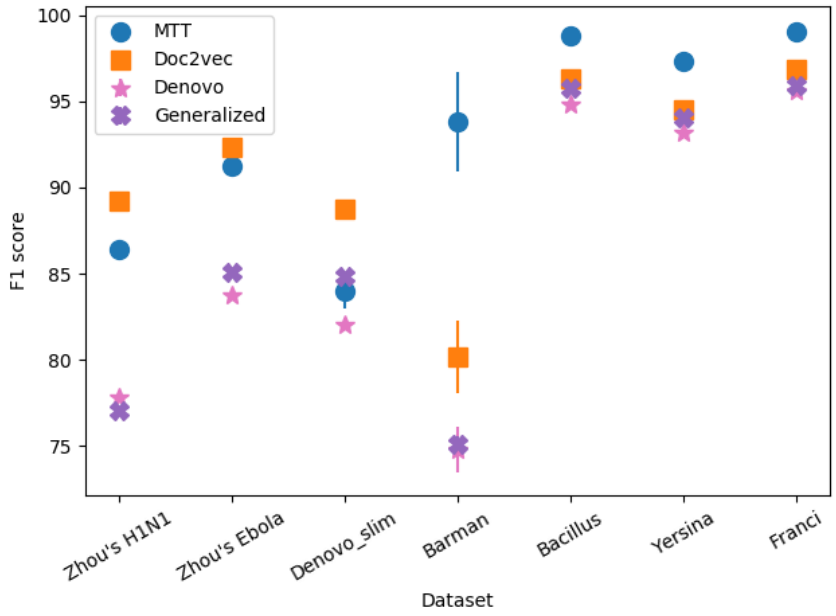


FIGURE 7.2: MTT vs. state-of-the-art methods on small testing datasets.

7.5.1 Comparison with methods employing hand-crafted features

GENERALIZED [246] and DENOVO [58] are the two traditional methods relying on hand-crafted features extracted from the protein sequences. The number of hand-crafted features employed by DENOVO and GENERALIZED are 686 and 1,175, respectively. They both employ SVM for the classification task. Since SVM scales quadratically with the number of data points, DENOVO and GENERALIZED are not scalable to larger datasets.

Figure 7.2 presents their comparison between MTT on small testing datasets. Detailed scores are given in Table 7.3. The performance gains are statistically significant with a p-value of 0.05. Results from the two-tailed t-test [103, 186] support that MTT significantly outperforms DENOVO in all benchmarked datasets with a confidence score of at least 95%. Compared with GENERALIZED, MTT has higher performance in six out of seven datasets (except DENOVO_SLIM). The difference is the most significant on the BARMAN, ZHOU'S H1N1, and ZHOU'S EBOLA datasets. On DENOVO_SLIM dataset, MTT's F1 score is lower than GENERALIZED and only 2% higher than DENOVO. This is expected since DENOVO_SLIM is a specialized dataset favoring methods using local sequence motif features, which are exploited by DENOVO and GENERALIZED.

HYBRID is one recently proposed, deep learning-based method. Despite that, the input features are still manually extracted from the protein sequence. Since the code is not publicly available, we only have the AUC score corresponding to the ZHOU'S H1N1 dataset, which is also taken from the original paper as listed in table 7.3. Compared with HYBRID, MTT has higher AUC score. Though comparison on the AUC for one dataset does not bring much insight, we include this method here for completeness.

TABLE 7.3: MTT vs. methods based on hand-crafted features.

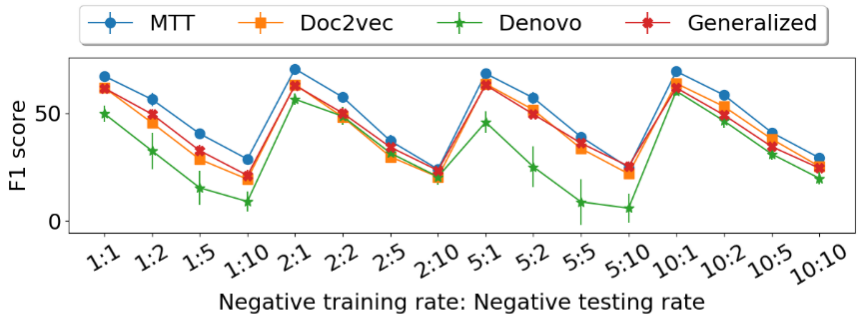
DATASET	MODEL	AUC	AP	PRE	REC	F1
ZHOU'S H1N1	DENOVO	0.8656	0.8619	77.75	77.95	77.85
	GENERALIZED	0.8600	0.8606	76.96	77.17	77.06
	HYBRID	0.937	-	-	-	-
	MTT	**0.9461**	**0.9589**	**86.28**	**86.51**	**86.40**
ZHOU'S EBOLA	DENOVO	0.8864	0.8366	83.44	84.00	83.72
	GENERALIZED	0.9154	0.9078	84.77	85.33	85.05
	MTT	**0.9680**	**0.9766**	**90.93**	**91.53**	**91.23**
DENOVO_SLIM	DENOVO	0.8701	0.8631	81.92	82.12	82.02
	GENERALIZED	0.8891	0.8851	**84.74**	**84.94**	**84.84**
	MTT	**0.9221**	**0.9324**	83.92	84.12	84.02
BARMAN	DENOVO	0.8217	0.8415	74.60	74.98	74.79
	GENERALIZED	0.8214	0.8458	74.90	75.27	75.08
	MTT	**0.9804**	**0.9802**	**93.53**	**94.05**	**93.79**
BACILLUS	DENOVO	0.9843	0.9650	94.80	94.83	94.83
	GENERALIZED	0.9833	0.9668	95.75	95.78	95.76
	MTT	**0.9997**	**0.9992**	**98.75**	**98.78**	**98.76**
YERSINA	DENOVO	0.9712	0.9302	93.14	93.16	93.15
	GENERALIZED	0.9758	0.9362	94.01	94.03	94.02
	MTT	**0.9988**	**0.9971**	**97.32**	**97.34**	**97.32**
FRANCI	DENOVO	0.9782	0.9584	95.55	95.62	95.58
	GENERALIZED	0.9799	0.9565	95.84	95.91	95.88
	MTT	**0.9998**	**0.9996**	**98.95**	**99.03**	**98.99**

7.5.2 Comparison with sequence embedding based methods

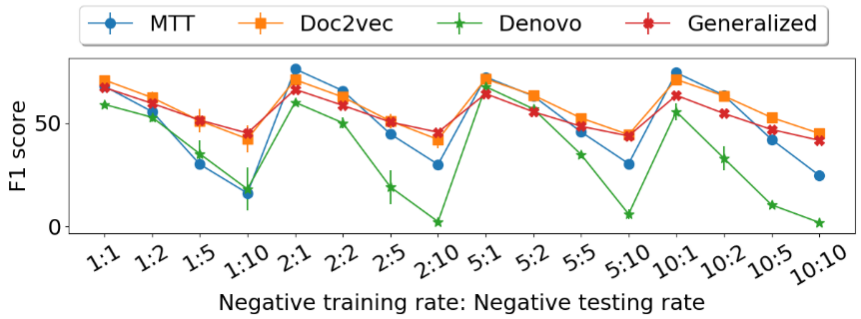
DOC2VEC and MOTIFTRANSFORMER are state-of-the-art methods based on sequence embeddings or representations. DOC2VEC utilizes the embeddings learned from the extracted k-mer features while MTT and MOTIFTRANSFORMER employ the embedding directly learned from the amino acid sequences. In addition, MTT is a multitask-based approach that incorporates additional information on human protein-protein interaction into the learning process.

Figure 7.3 shows a comparison in F1 score of MTT and DOC2VEC over all benchmarked datasets. Detailed scores are presented in Table 7.4. Since the code for the MOTIFTRANSFORMER model is not publicly available, we only have the corresponding results available for the ZHOU'S H1N1 and ZHOU'S EBOLA datasets, which are also taken from the original paper. '-' denotes the score is not available. Compared with MOTIFTRANSFORMER, MTT has a slightly worse F1 score on ZHOU'S H1N1 and significantly better F1 score on ZHOU'S EBOLA datasets.

Comparison with DOC2VEC. MTT out-performs DOC2VEC in 5 out of 9 benchmark datasets, and the performance gap is statistically significant with a p-value smaller than 0.05. MTT is significantly better than DOC2VEC on the



(a) Novel Ebola



(b) Novel H1N1

Figure 7.3: MTT vs. state-of-the-art methods on the Novel Ebola and Novel H1N1 datasets over different combinations of negative training and testing sets.

Table 7.4: MTT vs. embedding-based methods

Dataset	Model	AUC	AP	Pre	Rec	F1
Zhou's H1N1	doc2vec	**0.9601**	**0.9674**	**89.04**	**89.34**	**89.19**
	MotifTransformer	0.945	—	-	-	86.50
	MTT	0.9461	0.9589	86.28	86.51	86.40
Zhou's Ebola	Doc2vec	**0.9781**	**0.9832**	**91.99**	**92.67**	**92.33**
	MotifTransformer	0.968	—	-	-	89.6
	MTT	0.9680	0.9766	90.93	91.53	91.23
Denovo_slim	doc2vec	**0.9644**	**0.9681**	**88.60**	**88.87**	**88.73**
	MTT	0.9221	0.9324	83.92	84.12	84.02
Barman	doc2vec	0.8671	0.8922	79.95	80.37	80.16
	MTT	**0.9804**	**0.9802**	**93.53**	**94.05**	**93.79**
Bacillus	doc2vec	0.9900	0.9739	96.29	96.32	96.31
	MTT	**0.9997**	**0.9992**	**98.75**	**98.78**	**98.76**
Yersina	doc2vec	0.9814	0.9510	94.50	94.52	94.51
	MTT	**0.9988**	**0.9971**	**97.32**	**97.34**	**97.32**
Franci	doc2vec	0.9878	0.9606	96.77	96.84	96.81
	MTT	**0.9998**	**0.9996**	**98.95**	**99.03**	**98.99**

NOVEL EBOLA dataset, while on the NOVEL H1N1 dataset, the reverse holds true. DOC2VEC outperforms MTT in three testing datasets whose negative samples were drawn from a sequence dissimilarity method. We also note that these datasets might be biased since in the ideal testing scenario, we do not have knowledge about the set of human proteins that interacted with the virus. Therefore, such dissimilarity-based negative sampling is infeasible.

7.5.3 Comparison with a method that utilizes protein domain information

BARMAN features set is constructed from the domain-domain association and the hand-crafted feature extracted from the protein sequences. Since the protein domain information is not available for all viral proteins, the BARMAN method has restricted application. A comparison between BARMAN and MTT is presented in table 7.5. Due to data and code availability, we only have the results for the BARMAN model on one dataset. From reported results, we could clearly see that MTT outperforms its competitor for a large margin in all available metrics.

TABLE 7.5: MTT vs. BARMAN- a method that utilizes protein domain information.

MODEL	AUC	AP	PRE	REC	F1
BARMAN	0.7300	–	–	67.00	69.41
MTT	0.9804	0.9802	93.53	94.05	93.79

7.5.4 Comparison with methods that used GO, taxonomy and phenotype information

DEEPVIRAL exploited the disease phenotypes, the viral taxonomies, and proteins' GO annotation to enrich its protein embeddings. Table 7.6 presents a comparison between MTT and DEEPVIRAL on the four datasets released by DEEPVIRAL's authors. The reported results on each dataset are the average after five experimental runs for DEEPVIRAL and ten experimental runs for MTT. Results from the pair-wise t-test indicate that MTT is significantly better than DEEPVIRAL on three out of four datasets. In addition, we observe MTT and STT significantly supersede their competitor regarding the averaged F1 score. The gain is more significant on smaller datasets (644788 and 333761)

7.5.5 Ablation Studies

We compare our method with two of its simpler variants: the STT and the NAIVE BASELINE baseline models. STT is the MTT model without the human PPI prediction task. NAIVE BASELINE concatenates the learned embeddings for the virus and human proteins to form the input to a Logistic Regression model. Figure 7.4 presents a comparison between the F1 score of MTT

TABLE 7.6: MTT vs. DEEPVIRAL.

DATASET	MODEL	AUC	AP	PRE	REC	F1
2697049	DEEPVIRAL	0.7288	0.0015	0.07	0.07	0.07
	MTT	**0.7566**	**0.0021**	**0.97**	**0.97**	**0.97**
333761	DEEPVIRAL	0.8009	0.0147	1.72	1.72	1.72
	MTT	**0.8160**	**0.0262**	**6.35**	**6.35**	**6.35**
2043570	DEEPVIRAL	**0.7708**	**0.0116**	0.52	0.52	0.52
	MTT	0.6956	0.0096	**1.89**	**1.91**	**1.90**
644788	DEEPVIRAL	0.9325	**0.0357**	**3.70**	**3.70**	3.70
	MTT	**0.9537**	0.0302	3.54	**22.04**	**5.46**

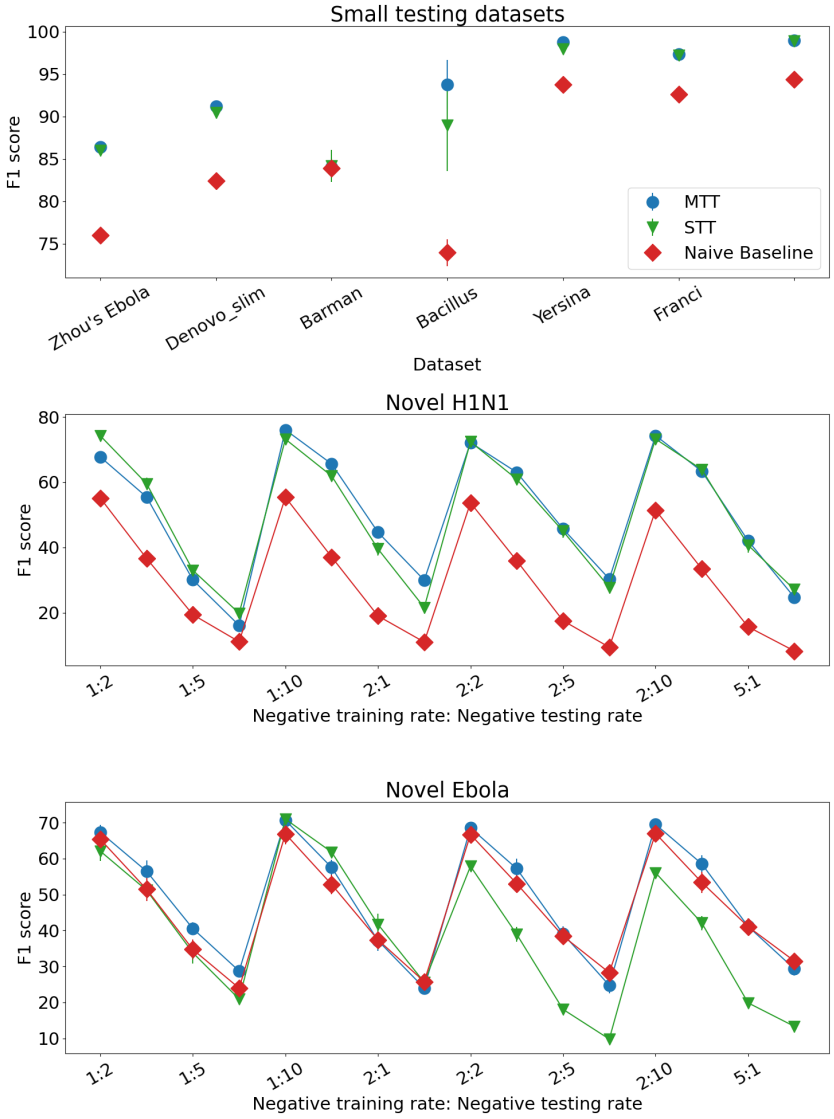


FIGURE 7.4: Ablation study on benchmarked datasets.

and its variants on our benchmarked datasets. Table 7.7 show all reported

scores over all datasets. The reported results are average after 10 runs. Results from pair-wise t-test indicate that MTT is significantly better than STT in five out of nine benchmarked and the four DEEPVIRAL datasets with a p-value smaller than 0.05. While in the remaining four datasets, the difference is not statistically significant. This confirms that the learned patterns from the human PPI network bring additional benefits to the virus-human PPI prediction task.

TABLE 7.7: Ablation study detailed results.

DATASET	MODEL	AUC	AP	PRE	REC	F1
H1N1	NAIVE BASELINE	0.8310	0.8003	75.92	76.12	76.02
	STT	0.9472	0.9590	85.86	86.09	85.98
	MTT	**0.9461**	**0.9589**	**86.28**	**86.51**	**86.40**
EBOLA	NAIVE BASELINE	0.8876	0.8665	82.12	82.67	82.39
	STT	0.9655	0.9749	90.13	90.73	90.43
	MTT	**0.9680**	**0.9766**	90.93	91.53	91.23
DENOVO_SLIM	NAIVE BASELINE	0.8843	0.8673	83.80	84.00	83.90
	STT	0.9207	0.9343	84.04	84.24	84.14
	MTT	**0.9221**	**0.9324**	**83.92**	**84.12**	**84.02**
BARMAN'S	NAIVE BASELINE	0.8084	0.8198	73.75	74.11	73.93
	MTT	0.9804	0.9802	93.53	94.05	93.79
	STT	0.9801	0.9802	93.83	94.29	94.06
BACILLUS	NAIVE BASELINE	0.9842	0.9619	93.75	93.78	93.77
	STT	0.9995	0.9986	97.93	97.96	97.95
	MTT	**0.9997**	**0.9992**	**98.75**	**98.78**	**98.76**
YERSINA	NAIVE BASELINE	0.9741	0.9277	92.61	92.64	92.63
	STT	**0.9987**	0.9970	97.18	97.30	97.24
	MTT	0.9988	**0.9971**	97.32	97.34	97.32
FRANCI	NAIVE BASELINE	0.9851	0.9680	94.36	94.43	94.39
	STT	0.9997	0.9993	98.84	98.92	98.88
	MTT	**0.9998**	**0.9996**	98.95	**99.03**	**98.99**
2697049	NAIVE BASELINE	0.5686	0.0010	0	0	0
	STT	0.7457	0.0017	0.07	0.07	0.07
	MTT	**0.7566**	**0.0021**	**0.97**	**0.97**	**0.97**
333761	NAIVE BASELINE	0.7002	0.0110	3.55	3.56	3.55
	STT	0.8114	0.0213	4.72	4.72	4.72
	MTT	0.8160	0.0262	**6.35**	6.35	**6.35**
2043570	NAIVE BASELINE	0.6624	0.0076	0.32	0.32	0.32
	STT	0.6706	0.0087	1.11	**3.01**	1.46
	MTT	**0.6956**	**0.0096**	**1.89**	1.91	**1.90**
644788	NAIVE BASELINE	0.8410	0.0089	1.82	1.85	1.83
	STT	**0.9705**	**0.0459**	3.97	9.26	4.65
	MTT	0.9537	0.0302	3.54	**22.04**	5.46

Compare with NAIVE BASELINE, MTT wins in eight out of nine benchmarked and the four DEEPVIRAL datasets. On the remaining dataset (NOVEL H1N1), the difference is not statistically different. STT significantly outperforms NAIVE BASELINE in eight out of nine datasets. This claims the effectiveness of our chosen architecture.

7.6 Case study for SARS-CoV-2 binding prediction

The virus binding to cells or the interaction between viral attachment proteins and host cell receptors is the first and decisive step in the virus replication cycle. Identifying the host receptor(s) for a particular virus is often fundamental in unveiling the virus pathogenesis and its species tropism.

Here we present a case study for detecting the human protein binding partners for SARS-COV-2. Our virus-human PPI dataset is retrieved from the InAct Molecular Interaction database [106] (the latest update is 07.05.2021). We retrieve the protein sequences from Uniprot [40]. In the next section, we describe the construction of the training and testing dataset to predict SARS-COV-2 binding partners.

7.6.1 Training, Validation and Test Sets for Virus-Human PPI

The statistics for our SARS-COV-2 binding prediction dataset are presented in table 7.8. We construct the corresponding datasets as follows.

Training Set. As positive interaction samples, we include in the training data only *direct* interactions between the human proteins and any virus except the SARS-COV and SARS-COV-2. *Direct* interaction requires two proteins to directly bind to each other, i.e. without an additional bridging protein. Moreover, the interacting human protein should be on the cell surface. Without loss of generality, we perform our search for the binding receptor on the set of all human proteins that have a *KNOWN direct interaction* with any virus and *locate* to the cell surface. Our surface human protein list consists of all reviewed Uniprot proteins that meet at least one of the following criteria: (i) appears in the human surfacetome [12] list or (ii) has at least one of the following GO annotations [7, 20]:{*CC-plasma membrane, CC-cell junction*}.

The negative samples for training data contain *indirect* (interactions that are not marked as direct in the database) between the human proteins and any virus except SARS-COV and SARS-COV-2. The *indirect* interactions can be a physical association (two proteins are detected in the same protein complex at the same point of time) or an association in which two proteins that may participate in the formation of one or more physical complexes without additional evidence whether the proteins are directly binding to specific members of such a complex).

Validation and Test Sets. As established in studies [84, 194, 241], angiotensin-converting enzyme 2 (ACE2) is the human receptor for both SARS-COV [128] and SARS-COV-2 viruses [84]. The positive validation and testing set consist of interaction between the known human receptor (ACE2) and the

corresponding spike proteins of SARS-CoV and SARS-CoV-2, respectively. Our negative validation and testing set encapsulate of all possible combinations the two viral spike proteins and 52 human proteins that meet our filtering criteria.

7.6.2 The intra human PPI for the Side Task

Since we are interested in only the direct interaction between virus and human proteins, we also customize our intra human PPI training set. Our intra human PPI dataset is also retrieved from the InAct [106] database (the latest update is 07.05.2021). We retain only interactions between two human proteins that appear in the virus-human PPI dataset constructed above. The confidence scores are normalized into the [0,1] ranges. All confidence scores corresponding to "indirect" interactions are set to 0. In the end, our intra-human PPI training set consists of 96,458 interactions between 5,563 human proteins.

TABLE 7.8: The case study statistics.

		TRAINING		VALIDATION		TESTING		HUMAN PPI
$\|V^h\|$	$\|V^v\|$	$\|E^+\|$	$\|E^-\|$	$\|E^+\|$	$\|E^-\|$	$\|E^+\|$	$\|E^-\|$	$\|E\|$
5,563	834	554	17,418	1	51	1	51	96,459

7.6.3 Results

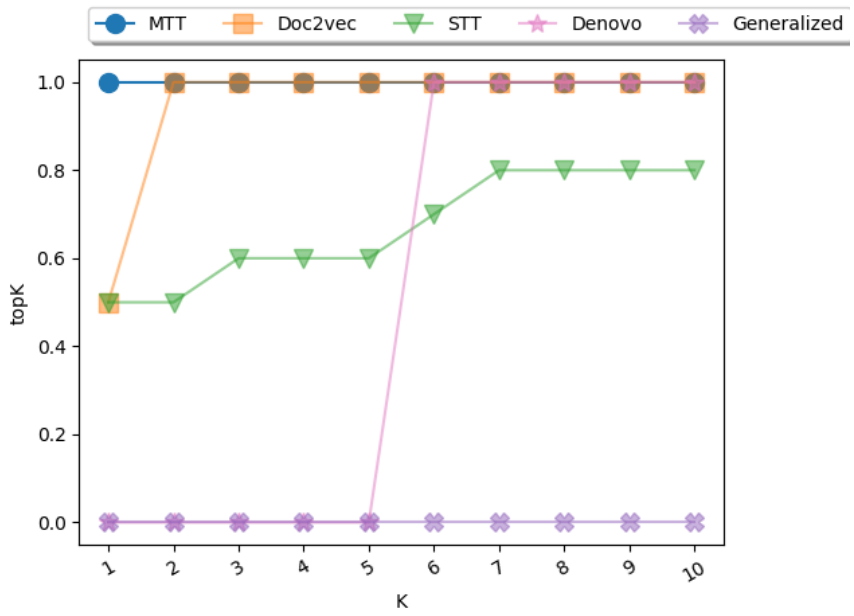


FIGURE 7.5: Case study results for benchmarked methods.

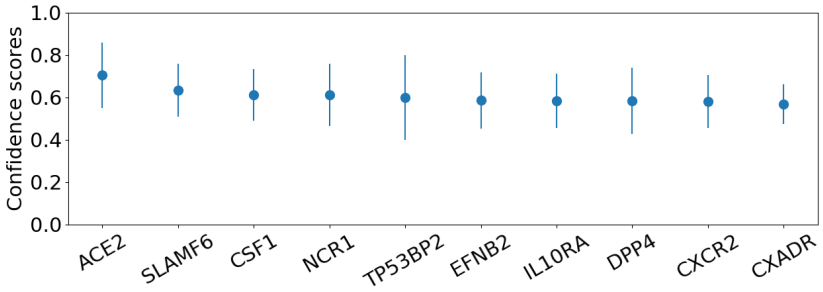


FIGURE 7.6: MTT's top 10 highest predictions in the virus binding case study.

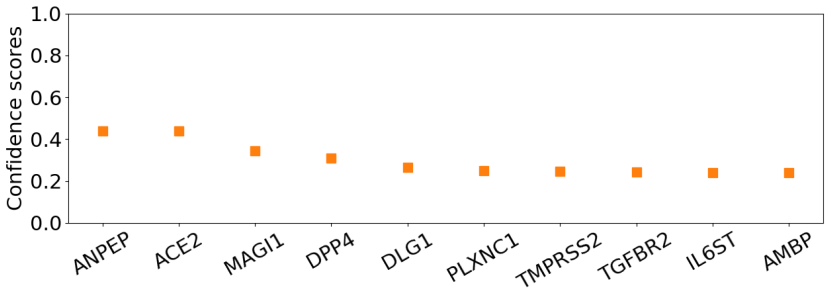


FIGURE 7.7: DOC2VEC's top 10 highest predictions in the virus binding case study.

Finally, we here evaluate the prediction methods on how effective they are in ranking human protein candidates for binding to an emerging virus envelope protein. Figure 7.5 presents the methods' performance after ten runs on the case study dataset. TopK is equal to 1 if the true human receptor appears in the top K proteins that correspond to the highest predicted scores by the model and is equal to 0 otherwise. The reported scores plotted in Figure 7.5 are the average after ten experimental runs with random initialization.

Using this method we find that ACE2, the only SARS-COV-2 receptor proven in *in vivo* and *in vitro* studies[9, 84, 227], consistently appears as the highest ranked prediction of MTT in each of the ten experimental runs. We observe a significant difference between the highest ranked performance of MTT and its competitors. The performance gain shown by MTT over STT is quite substantial after ten runs and supports the superiority of our multi-task framework. The next highest nine hits presented in both models have not been shown to interact with SARS-CoV-2 in *in vitro* studies. Interestingly, dipeptidyl peptidase 4 (DDP4), a receptor for another betacoronavirus MERS-CoV [221] also scored highly in the MTT method. However, although *in silico* analysis has speculated a possible interaction [209], it is yet to be shown experimentally. Similarly, the serine protease TMPRSS2, which is required for SARS-CoV-2 S protein priming during entry [84], appeared in position 7 using the Doc2vec model. Finally, aminopeptidase N (ANPEP) the receptor for the common cold coronavirus 229E appeared as first hit in the Doc2vec model [237].

In Figures 7.6 and 7.7, we plot the average confidence scores (corresponding to predicted interaction probability) corresponding to top 10 predictions

of MTT and DOC2VEC models. The dots represent the average confidence scores after 10 experimental runs while the lines represent the standard deviation. Specifically, the proteins are ranked based on the average (over 10 runs) confidence scores as predicted by the two models. While for MTT, the receptor ACE2 always occurs at the top of the list with average confidence score of more than 0.70 (which is more than 11% higher than the confidence score assigned to the second hit), DOC2VEC assigns it a score of less than 0.44 where ACE2 is ranked 2nd based on average scores. Moreover, there is negligible difference between the prediction scores for ACE2 and the first predicted hit ANPEP in case of DOC2VEC.

These results indicate that MTT can provide high-quality prediction results and can help biologists to restrict the search space for the virus interaction partner effectively. This case study showcases the effectiveness of our method in solving virus-human PPI prediction problem and aims to convince biologists of the potential application of our prediction framework.

7.7 Conclusion

We presented a thorough overview of state-of-the-art models and their limitations for the task of virus-human PPI prediction. Our proposed approach exploits powerful statistical protein representations derived from a corpus of around 24 Million protein sequences in a multitask framework. Noting the fact that virus proteins tend to mimic human proteins towards interacting with the host proteins, we use the prediction of human PPI as a side task to regularize our model and improve generalization. The comparison of our method with a variety of state-of-the-art models on several datasets showcase the superiority of our approach. Ablation study results suggest that the human PPI prediction side task brings additional benefits and helps boost the model performance. A case study on the interaction of the SARS-COV-2 virus spike protein and its human receptor indicates that our model can be used as an effective tool to reduce the search space for evaluating host protein candidates as interacting partners for emerging viruses. In future work, we will enhance our multitask approach by incorporating more domain information including structural protein prediction tools [99] as well as exploiting more complex multitask model architectures.

Chapter 8

Conclusion and future outlook

8.1 Conclusion

More and more data is becoming available in recent years and in the future. Data availability unveils unprecedented opportunities for machine learning models on biological problems. Yet biological data is biased and limited in quality and quantity. Those challenges lead to biased and non-generalizable models. In addition, data scarcity restricts the application of deep learning techniques that require a large amount of annotated data to train. This thesis focuses on machine learning models for two biological problems: the miRNA-disease association and virus-human protein-protein interaction prediction. To summarize, we make the following contributions.

8.1.1 Identification and analysis of existing systems' limitations

We start with identifying and experimentally analyzing some critical limitations of existing works. We pinpoint the data leakage problem that results in overestimating methods' performance and unfair comparison between models. Besides, we present some other issues related to existing systems' experimental setup and evaluation metrics. To address those limitations, we develop and release a consistent evaluation framework with the implementation of various similarity calculations, a consensus evaluation metric, and updates to the existing works' training workflow.

8.1.2 New model development

The second focus of the thesis is on new model development. The motivation for our work is that collecting and cleaning annotated training biological data for a particular machine learning problem is often expensive, time-consuming, and even unrealistic in some scenarios. We notice that besides the given small annotated data, other publicly available related information sources exist. Our strategy in such a context is to develop machine learning frameworks that can exploit such related knowledge. Yet simple concatenation does not work, especially when there are limited training samples. We try to provide the answers to three research questions: (i) which are the available information sources? (ii) how can we incorporate such information into

our systems? and (iii) concerning a particular source, how much should we add?

Joint learning architectures

To answer the above questions, for each biological problem, we identify the available related data and their corresponding biological rationale. We then propose joint learning models that can flexibly integrate multiple side information sources at different stages of the model building process. The integrated knowledge sources can be fused as the data to learn biologically rich statistical representations, as the materials to construct the input networks for representation learning, as the raw features, as the supervised signals to guide our data preprocessing and filtering module, or as the training data for the additional side tasks. We employ the architectures proposed for language modeling, multitask learning, graph representation learning, and feature selection to construct our joint learning systems. Combining information helps us to overcome the data scarcity issue. At the same time, such an approach offers global views of the target problems that enable machine learning models to generate more reliable predictions. Results from large-scale experiments claim that our proposed architectures acquire state-of-the-art performance on the selected biological problems.

Data preprocessing

Besides presenting the answers to the questions of which and how to integrate, we also come up with ways to control the quality and quantity of the incorporated data sources. Our data preprocessing pipelines encapsulate various processing techniques, from naive threshold filtering to more complicated approaches that employ field experts' domain knowledge. For example, concerning the virus receptor prediction task, we clean the added side information by selecting only the human protein-protein interactions that are marked as '*direct binding*' and are between two human proteins that show up at the *cell membrane*. Concerning the miRNA-disease association prediction, we propose a parameter-free learning module that is motivated by biological heuristics to enrich and filter the incorporated miRNA and disease-protein coding gene associations. Such effective data preprocessing strategies also contribute toward the success of our joint learning models.

8.1.3 Fair and comprehensive evaluation

The third focus area of the thesis is on fair and comprehensive comparison among benchmarked systems. As the given annotated data is scarce and biased, evaluating machine learning approaches on little-known or completely new entities is an important evaluation criterion that has been neglected in existing works. To address such an issue, we curate and release new datasets to enable the assessments under various testing scenarios, including those with realistic negative sample rates and those with little or completely new

entities. We believe such newly created datasets will enable large-scale experiments and, thus, facilitate future research and development in the field.

8.1.4 Support for end-users assessment and adoption

Our final effort concentrates on supporting field experts' assessment and adoption. In our work, we add realistic case studies in which we place no simplified assumptions on the potential candidate search space. Such newly added case studies offer the answers to challenging and practical questions like (i) what is the proposed method's performance on a completely new or little-known disease? (ii) how well the system can differentiate between true and false positives, given the noisy training data? (iii) how good the system is in identifying the human receptor for a completely new virus? or (iv) how much do our predictions agree with the results acquired from a survival analysis on publicly available patient data? In addition, we develop and release an easy-to-use web application that encapsulates all our model's generated predictions as well as the related biological knowledge to support field experts' assessment and future adoption.

Though the answers to the questions of which and how to incorporate the related information sources depend heavily on the learning tasks, we do believe that our design principles, the fair and comprehensive evaluation strategy, the realistic case studies, and the easy-to-use web application will be of independent interest to and would help facilitate future research in biomedical applications.

8.2 Future outlook

In our opinion, future works can go in one of the following four potential directions.

8.2.1 New model development or new application for existing models

One can develop or adapt existing joint learning approaches for other supervised prediction problems. We believe that our models would benefit other prediction tasks where data scarcity and biased issues persist. For example, one can apply our proposed joint learning model to the miRNA-small molecule drug association prediction problem as discussed in Section 6.7. Similarly, one can adopt the existing multitask framework for virus-human protein-protein interaction prediction to tackle some of the realistic research questions in drug development like the virus human receptor prediction, predicting the human protein that helps the virus to replicate, or predicting the human protein that helps the virus to get out of the human cell, etc. At the same time, researchers can propose joint learning approaches for other biological problems where the related knowledge sources exist.

8.2.2 End-user experience enhancement

People might focus on end-user experience enhancement. Works focusing on *user experience enhancement* should provide a user-friendly interface like a portable application or a publicly available website. Such easy-to-use tools would enlarge the research impact by facilitating end-users assessments and adaptation. Besides, some of the nice-to-have features of the tool would include:

Automation of data/results filtering with different filtering criteria. This functionality would enable field experts to select only the subset of entities that they are interested in or take a closer look at the generated results to validate their hypothesis or assess the generated predictions based on their domain knowledge.

Support for hypothesis testing by integrating more related biological information like the miRNA tissue expression profile, miRNA chromosomal location, clinical disease phenotype, etc. Such a feature would enable users to draw biological insights about the generated predictions based on their expertise.

Comparison of the generated predictions from different models. As the performance of machine learning models varies given the input data, it could be the case that model **A** generates better prediction than model **B** on dataset $\mathcal{D}$, but model **B** can be better than model **A** on some specialized dataset $\mathcal{D}'$. Therefore, enabling the user to compare models' performance with different input datasets offers them the ability to choose the best model for their own data, thus, facilitating future research and increasing the reach and impact of existing machine learning models.

Possibility to train the model on the fly with user-customized data. Most publicly available models are only trained or fine-tuned on public databases. Nevertheless, data sharing is one of the critical issues in biological applications. Because of privacy and many other factors, field experts might want to have the model trained on their in-house or private data. Yet not everyone is familiar with coding and machine learning system training and testing. Offering the model to be trained/tested on users' customized data with just a few clicks enable non-expert users to adapt existing models to their own problem.

Allowing configurable model parameters. This functionality serves as an add-on to the above feature by enabling advanced users to find the hyperparameter set that is best fitted to their customized dataset.

8.2.3 Input data or feature enhancement

One can also focus on input data or feature enhancements. In what follows, researchers could propose the incorporation of additional data sources or a more complex learning architecture. For example, for the virus-human protein-protein interaction prediction problem, one future work would be to integrate additional relevant domain information like the experimentally

verified or predicted protein structural information (from computation prediction tool like [99]) or to exploit more complex multitask learning architectures. In addition, since data quality significantly affects the learning model's outcome, the other fruitful approach would be to focus on effective data filtering or expert-driven negative sample selection strategies. Properly cleaned data with 'true' negative samples help the model to learn informative patterns and, thus, help it to improve the prediction performance.

8.2.4 Model explainability

One future direction would be to focus on explainability or models that can generate explanations. As the majority of machine learning systems focus on supporting end-user decision-making processes, generated explanations that can help in model debugging and end-user convincing is one important aspect in future machine learning applications. One can employ post-hoc explanation techniques [66] to generate instance-level explanations or utilize association rule-based models to extract the set of rules/reasons that lead to the model's decision. Nevertheless, domain expertise will be required to translate these explanations into biological rationales.

Bibliography

[1] Khan Academy. *https://www.khanacademy.org/science/biology/gene-regulation/gene-regulation-in-eukaryotes/a/regulation-after-transcription*. 2013.

[2] *Adaboost classifier*. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html`.

[3] Ethan C Alley et al. "Unified rational protein engineering with sequence-based deep representation learning". In: *Nature methods* 16.12 (2019), pp. 1315–1322.

[4] Diego Alonso-Lopez et al. "APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks". In: *Nucleic acids research* 44.W1 (2016), W529–W535.

[5] Mais G Ammari et al. "HPIDB 2.0: a curated database for host–pathogen interactions". In: *Database* 2016 (2016).

[6] Bruno Aranda et al. "PSICQUIC and PSISCORE: accessing and scoring molecular interactions". In: *Nature methods* 8.7 (2011), pp. 528–529.

[7] Michael Ashburner et al. "Gene ontology: tool for the unification of biology". In: *Nature genetics* 25.1 (2000), pp. 25–29.

[8] SM Bailer and J Haas. "Connecting viral with cellular interactomes". In: *Current opinion in microbiology* 12.4 (2009), pp. 453–459.

[9] Linlin Bao et al. "The pathogenicity of SARS-CoV-2 in hACE2 transgenic mice". In: *Nature* 583.7818 (2020), pp. 830–833.

[10] Ranjan Kumar Barman, Sudipto Saha, and Santasabuj Das. "Prediction of interactions between viral and host proteins using supervised machine learning methods". In: *PloS one* 9.11 (2014), e112034.

[11] Abdul Hannan Basit et al. "Training host-pathogen protein–protein interaction predictors". In: *Journal of bioinformatics and computational biology* 16.04 (2018), p. 1850014.

[12] Damaris Bausch-Fluck et al. "A mass spectrometric-derived cell surface protein atlas". In: *PloS one* 10.4 (2015), e0121314.

[13] Pierre M Jean Beltran, Katelyn C Cook, and Ileana M Cristea. "Exploring and Exploiting Proteome Organization during Viral Infection". In: *Journal of Virology* 91.18 (2017), e00268–17.

[14] Visar Berisha et al. "Digital medicine and the curse of dimensionality". In: *NPJ digital medicine* 4.1 (2021), pp. 1–8.

[15] Sanmitra Bhattacharya, Viet Ha-Thuc, and Padmini Srinivasan. "MeSH: a window into full text for document summarization". In: *Bioinformatics* 27.13 (2011), pp. i120–i128.

[16] Surojit Biswas. "Principles of Machine Learning-Guided Protein Engineering". PhD thesis. 2020.

[17] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[18] Yimei Cai et al. "A brief review on the mechanisms of miRNA regulation". In: *Genomics, proteomics & bioinformatics* 7.4 (2009), pp. 147–154.

[19] Alberto Calderone, Luana Licata, and Gianni Cesareni. "VirusMentha: a new resource for virus-host protein interactions". In: *Nucleic acids research* 43.D1 (2015), pp. D588–D592.

[20] Seth Carbon et al. "The Gene Ontology resource: enriching a GOld mine". In: *Nucleic Acids Research* 49.D1 (2021), pp. D325–D334.

[21] Rich Caruana. "Multitask learning". In: *Machine learning* 28.1 (1997), pp. 41–75.

[22] Fulvia Ceccarelli et al. "Prediction of chronic damage in systemic lupus erythematosus by using machine-learning models". In: *PLoS One* 12.3 (2017), e0174200.

[23] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. "Semi-supervised learning. 2006". In: *Cambridge, Massachusettes: The MIT Press View Article* 2 (2006).

[24] Andrew Chatr-Aryamontri et al. "VirusMINT: a viral protein interaction database". In: *Nucleic acids research* 37.suppl_1 (2009), pp. D669–D673.

[25] Hailin Chen and Zuping Zhang. "A miRNA-driven inference model to construct potential drug-disease associations for drug repositioning". In: *BioMed research international* 2015 (2015).

[26] Hailin Chen and Zuping Zhang. "Prediction of drug-disease associations for drug repositioning through drug-miRNA-disease heterogeneous network". In: *IEEE Access* 6 (2018), pp. 45281–45287.

[27] Hailin Chen and Zuping Zhang. "Similarity-based methods for potential human microRNA-disease association prediction". In: *BMC medical genomics* 6.1 (2013), p. 12.

[28] Hailin Chen, Zuping Zhang, and Wei Peng. "miRDDCR: a miRNA-based method to comprehensively infer drug-disease causal relationships". In: *Scientific reports* 7.1 (2017), pp. 1–9.

[29] Kuan-Hsi Chen, Tsai-Feng Wang, and Yuh-Jyh Hu. "Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme". In: *BMC bioinformatics* 20.1 (2019), pp. 1–17.

[30] Xing Chen, Ming-Xi Liu, and Gui-Ying Yan. "RWRMDA: predicting novel human microRNA–disease associations". In: *Molecular BioSystems* 8.10 (2012), pp. 2792–2798.

[31] Xing Chen, Lian-Gang Sun, and Yan Zhao. "NCMCMDA: miRNA–disease association prediction through neighborhood constraint matrix completion". en. In: *Briefings in Bioinformatics* 22.1 (Jan. 2021), pp. 485–496. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbz159. URL: https://academic.oup.com/bib/article/22/1/485/5685754 (visited on 06/10/2022).

[32] Xing Chen and Gui-Ying Yan. "Semi-supervised learning for potential human microRNA-disease associations inference". In: *Scientific reports* 4.1 (2014), p. 5501.

[33] Xing Chen, Chi-Chi Zhu, and Jun Yin. "Ensemble of decision tree reveals potential miRNA-disease associations". eng. In: *PLoS computational biology* 15.7 (July 2019), e1007209. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007209.

[34] Xing Chen et al. "Deep-belief network for predicting potential miRNA-disease associations". en. In: *Briefings in Bioinformatics* 22.3 (May 2021), bbaa186. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbaa186. URL: https://academic.oup.com/bib/article/doi/10.1093/bib/bbaa186/5898648 (visited on 06/10/2022).

[35] Xing Chen et al. "HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction". In: *Oncotarget* 7.40 (2016), pp. 65257–65269.

[36] Xing Chen et al. "MicroRNA-small molecule association identification: from experimental results to computational models". In: *Briefings in bioinformatics* 21.1 (2020), pp. 47–61.

[37] Xing Chen et al. "Predicting miRNA–disease association based on inductive matrix completion". In: *Bioinformatics* 34.24 (2018), pp. 4256–4265.

[38] Xing Chen et al. "WBSMDA: within and between score for MiRNA-disease association prediction". In: *Scientific reports* 6.1 (2016), p. 21106.

[39] V Gregory Chinchar. "Replication of viruses". In: *Encyclopedia of Virology* (1999), p. 1471.

[40] UniProt Consortium. "UniProt: a hub for protein information". In: *Nucleic acids research* 43.D1 (2015), pp. D204–D212.

[41] Guangyu Cui, Chao Fang, and Kyungsook Han. "Prediction of protein-protein interactions between viruses and human by an SVM model". In: *BMC bioinformatics*. Vol. 13. 7. Springer. 2012, pp. 1–10.

[42] Cameron Davidson-Pilon. "lifelines: survival analysis in Python". In: *Journal of Open Source Software* 4.40 (2019), p. 1317.

[43] *Decision Tree classifier*. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.

[44] Noemi Del Toro et al. "The IntAct database: efficient access to fine-grained molecular interaction data". In: *Nucleic acids research* 50.D1 (2022), pp. D648–D653.

[45] Lei Deng, Jiaojiao Zhao, and Jingpu Zhang. "Predict the Protein-protein Interaction between Virus and Host through Hybrid Deep Neural Network". In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2020, pp. 11–16.

[46] Lopamudra Dey, Sanjay Chakraborty, and Anirban Mukhopadhyay. "Machine learning techniques for sequence-based prediction of viral–host interactions between SARS-CoV-2 and human proteins". In: *Biomedical journal* 43.5 (2020), pp. 438–450.

[47] Kevin Dick et al. "Pipe4: Fast ppi predictor for comprehensive inter- and cross-species interactomes". In: *Scientific reports* 10.1 (2020), pp. 1–15.

[48] Francesca Diella et al. "Understanding eukaryotic linear motifs and their role in cell signaling and regulation". In: *Front Biosci* 13.6580 (2008), p. 603.

[49] Yulian Ding et al. "Variational graph auto-encoders for miRNA-disease association prediction". In: *Methods* 192 (2021), pp. 25–34.

[50] Ngan Thi Dong and Megha Khosla. "Revisiting feature selection with data complexity". In: *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE. 2020, pp. 211–216.

[51] Thi Ngan Dong and Megha Khosla. "Towards a consistent evaluation of miRNA-disease association prediction models". In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2020, pp. 1835–1842.

[52] Thi Ngan Dong, Stefanie Mucke, and Megha Khosla. "MuCoMiD: A Multitask graph Convolutional Learning Framework for miRNA-Disease Association Prediction". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2022), pp. 1–1. DOI: 10.1109/TCBB.2022.3176456.

[53] Thi Ngan Dong et al. "A Message Passing framework with Multiple data integration for miRNA-Disease association prediction". In: *Scientific Reports* (2022), p. 16259.

[54] Thi Ngan Dong et al. "A multitask transfer learning framework for the prediction of virus-human protein–protein interactions". In: *BMC bioinformatics* 22.1 (2021), pp. 1–24.

[55] Yadong Dong et al. "EPMDA: edge perturbation based method for miRNA-disease association prediction". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 17.6 (2019), pp. 2170–2175.

[56] *Drug ATC code*. https://www.whocc.no/atc_ddd_index/.

[57] Lynn B Dustin et al. "Hepatitis C virus: life cycle in cells, infection and host response, and analysis of molecular markers influencing the outcome of infection and response to therapy". In: *Clinical microbiology and infection* 22.10 (2016), pp. 826–832.

[58] Fatma-Elzahraa Eid, Mahmoud ElHefnawi, and Lenwood S Heath. "DeNovo: virus-host sequence-based protein–protein interaction prediction". In: *Bioinformatics* 32.8 (2016), pp. 1144–1150.

[59] Nagwa Elaraby, Sherif Barakat, and Amira Rezk. "A conditional GAN-based approach for enhancing transfer learning performance in few-shot HCR tasks". In: *Scientific Reports* 12.1 (2022), pp. 1–18.

[60] Terry S Elton, Sarah E Sansom, and Mickey M Martin. "Trisomy-21 gene dosage over-expression of miRNAs results in the haploinsufficiency of specific target proteins". In: *RNA biology* 7.5 (2010), pp. 540–547.

[61] Antonio Fabregat et al. "Reactome pathway analysis: a high-performance in-memory approach". In: *BMC bioinformatics* 18.1 (2017), pp. 1–9.

[62] Antonio Fabregat et al. "The reactome pathway knowledgebase". In: *Nucleic acids research* 46.D1 (2018), pp. D649–D655.

[63] Scott Federhen. "The NCBI taxonomy database". In: *Nucleic acids research* 40.D1 (2012), pp. D136–D143.

[64] William A Figgett et al. "Machine learning applied to whole-blood RNA-sequencing data uncovers distinct subsets of patients with systemic lupus erythematosus". In: *Clinical & translational immunology* 8.12 (2019), e01093.

[65] Guodong Fu et al. "MicroRNAs in human placental development and pregnancy complications". In: *International journal of molecular sciences* 14.3 (2013), pp. 5519–5544.

[66] Thorben Funke et al. "Zorro: Valid, Sparse, and Stable Explanations in Graph Neural Networks". In: *IEEE Transactions on Knowledge and Data Engineering* (2022), pp. 1–12. DOI: 10.1109/TKDE.2022.3201170.

[67] *Gaussian Naive Bayes classifier*. https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html.

[68] *GDC Data Portal*. https://gdc.cancer.gov/.

[69] *GDC Data Transfer Tool*. https://gdc.cancer.gov/access-data/gdc-data-transfer-tool.

[70] Gisa Gerold, Janina Bruening, and Thomas Pietschmann. "Decoding protein networks during virus entry by quantitative proteomics". In: *Virus research* 218 (2016), pp. 25–39.

[71] Gisa Gerold et al. "Protein Interactions during the Flavivirus and Hepacivirus Life Cycle". In: *Molecular and Cellular Proteomics* 16.4 suppl 1 (2017), S75–S91.

[72] Coryandar Gilvary et al. "A machine learning and network framework to discover new indications for small molecules". In: *PLoS computational biology* 16.8 (2020), e1008098.

[73] Leonid Gitlin et al. "Rapid evolution of virus sequences in intrinsically disordered protein regions". In: *PLoS pathogens* 10.12 (2014), e1004529.

[74] Yury E Glazyrin et al. "Proteomics-based machine learning approach as an alternative to conventional biomarkers for differential diagnosis of chronic kidney diseases". In: *International Journal of Molecular Sciences* 21.13 (2020), p. 4802.

[75] Yuchong Gong et al. "A network embedding-based multiple information integration method for the MiRNA-disease association prediction". In: *BMC bioinformatics* 20.1 (2019), p. 468.

[76] Assaf Gottlieb et al. "PREDICT: a method for inferring novel drug indications with application to personalized medicine". In: *Molecular systems biology* 7.1 (2011), p. 496.

[77] Todd M Greco and Ileana M Cristea. "Proteomics tracing the footsteps of infectious disease". In: *Molecular & Cellular Proteomics* 16.4 (2017), S5–S14.

[78] Na-Na Guan et al. "Prediction of potential small molecule-associated microRNAs using graphlet interaction". In: *Frontiers in pharmacology* 9 (2018), p. 1152.

[79] Emine Guven-Maiorov et al. "Interface-based structural prediction of novel host-pathogen interactions". In: *Computational Methods in Protein Evolution 2019*. Springer, pp. 317–335.

[80] David Harrington. "Linear rank tests in survival analysis". In: *Encyclopedia of biostatistics* (2005). DOI: 10.1002/0470011815.b2a11047.

[81] G Traver Hart, Arun K Ramani, and Edward M Marcotte. "How complete are current yeast and human protein-interaction networks?" In: *Genome biology* 7.11 (2006), pp. 1–9.

[82] Masahiro Hattori et al. "Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways". In: *Journal of the American Chemical Society* 125.39 (2003), pp. 11853–11865.

[83] C. Henley. *Foundations of Neuroscience*. Open textbook library. Michigan State University, 2021. URL: https://books.google.nl/books?id=rRCKzgEACAAJ.

[84] Markus Hoffmann et al. "SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor". In: *cell* 181.2 (2020), pp. 271–280.

[85] Alberta Hoi et al. "Algorithm for calculating high disease activity in SLE". In: *Rheumatology* 60.9 (2021), pp. 4291–4297.

[86] Hsi-Yuan Huang et al. "miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database". In: *Nucleic acids research* 48.D1 (2020), pp. D148–D154.

[87] Zhou Huang et al. "Benchmark of computational methods for predicting microRNA-disease associations". In: *Genome biology* 20.1 (2019), pp. 1–13.

[88] Zhou Huang et al. "HMDD v3. 0: a database for experimentally supported human microRNA–disease associations". In: *Nucleic acids research* 47.D1 (2019), pp. D1013–D1017.

[89] Sohyun Hwang et al. "HumanNet v2: human gene networks for disease research". In: *Nucleic acids research* 47.D1 (2019), pp. D573–D580.

[90] Asmaa Ibrahim et al. "Artificial intelligence in digital breast pathology: techniques and applications". In: *The Breast* 49 (2020), pp. 267–273.

[91] Salma Jamal, Vinita Periwal, Vinod Scaria, et al. "Computational analysis and predictive modeling of small molecule modulators of microRNA". In: *Journal of cheminformatics* 4.1 (2012), p. 16.

[92] Pierre M Jean Beltran, Katelyn C Cook, and Ileana M Cristea. "Exploring and exploiting proteome organization during viral infection". In: *Journal of virology* 91.18 (2017), e00268–17.

[93] Bo-Ya Ji et al. "Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model". In: *Scientific Reports* 10.1 (2020), p. 6658.

[94] Fei Jiang et al. "Artificial intelligence in healthcare: past, present and future". In: *Stroke and vascular neurology* 2.4 (2017).

[95] Limin Jiang et al. "MDA-SKF: similarity kernel fusion for accurately discovering miRNA-disease association". In: *Frontiers in Genetics* 9 (2018), p. 618.

[96] Qinghua Jiang et al. "miR2Disease: a manually curated database for microRNA deregulation in human disease". In: *Nucleic acids research* 37.suppl_1 (2009), pp. D98–D104.

[97] Fangfang Jin et al. "Serum microRNA profiles serve as novel biomarkers for autoimmune diseases". In: *Frontiers in immunology* 9 (2018), p. 2381.

[98] April Jorge et al. "Identifying lupus patients in electronic health records: Development and validation of machine learning algorithms and application of rule-based algorithms". In: *Seminars in arthritis and rheumatism*. Vol. 49. 1. Elsevier. 2019, pp. 84–90.

[99] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* (2021), pp. 1–11.

[100] Alexander Junge et al. "RAIN: RNA–protein association and interaction networks". In: *Database* 2017 (2017).

[101] *K-nearest neighbor classifier*. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html.

[102] Bogumił Kaczkowski et al. "Structural profiles of human miRNA families from pairwise clustering". In: *Bioinformatics* 25.3 (2009), pp. 291–294.

[103] Karen Kafadar. "Handbook of parametric and nonparametric statistical procedures". In: *The American Statistician* 51.4 (1997), p. 374.

[104] Brian Kegerreis et al. "Machine learning approaches to predict lupus disease activity from gene expression data". In: *Scientific reports* 9.1 (2019), pp. 1–12.

[105] Andreas Keller et al. "Toward the blood-borne miRNome of human diseases". In: *Nature methods* 8.10 (2011), pp. 841–843.

[106] Samuel Kerrien et al. "The IntAct molecular interaction database in 2012". In: *Nucleic acids research* 40.D1 (2012), pp. D841–D846.

[107] Byungmin Kim et al. "An improved method for predicting interactions between virus and human proteins". In: *Journal of bioinformatics and computational biology* 15.01 (2017), p. 1650024.

[108] V Narry Kim and Jin-Wu Nam. "Genomics of microRNA". In: *TRENDS in Genetics* 22.3 (2006), pp. 165–173.

[109] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *International Conference on Learning Representations (ICLR)* (2015).

[110] Adrienne Kline. *https://medium.com/tech-iiitg/multimodal-meta-learning-with-siamese-network-metric-space-method-a704e6017271*. 2022.

[111] Igor Kononenko. "Estimating attributes: analysis and extensions of RELIEF". In: *European conference on machine learning*. 1994, pp. 171–182.

[112] Mario Köppen. "The curse of dimensionality". In: *5th online world conference on soft computing in industrial applications (WSC5)*. Vol. 1. 2000, pp. 4–8.

[113] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. "miRBase: from microRNA sequences to function". In: *Nucleic acids research* 47.D1 (2019), pp. D155–D162.

[114] Ana Kozomara and Sam Griffiths-Jones. "miRBase: integrating microRNA annotation and deep-sequencing data". In: *Nucleic acids research* 39.suppl_1 (2010), pp. D152–D157.

[115] Ming-Che Kuo et al. "The role of noncoding RNAs in Parkinson's disease: biomarkers and associations with pathogenic pathways". In: *Journal of biomedical science* 28.1 (2021), p. 78.

[116] Twan van Laarhoven, Sander B Nabuurs, and Elena Marchiori. "Gaussian interaction profile kernels for predicting drug–target interaction". In: *Bioinformatics* 27.21 (2011), pp. 3036–3043.

[117] Wei Lan et al. "Predicting microRNA-disease associations based on improved microRNA and disease similarities". In: *IEEE/ACM transactions on computational biology and bioinformatics* 15.6 (2016), pp. 1774–1782.

[118] Jack Lanchantin et al. "Transfer Learning for Predicting Virus-Host Protein Interactions for Novel Virus Sequences". In: *bioRxiv* (2021), pp. 2020–12.

[119] Gorka Lasso et al. "A structure-informed atlas of human-virus interactions". In: *Cell* 178.6 (2019), pp. 1526–1541.

[120] Lisa Lasswitz et al. "Glycomics and proteomics approaches to investigate early adenovirus–host cell interactions". In: *Journal of molecular biology* 430.13 (2018), pp. 1863–1882.

[121] Ke W LC, R Ga, et al. "Serum Metabolomic Signatures Can Predict Subclinical Atherosclerosis in Patients With Systemic Lupus Erythematosus.[J]". In: *Arteriosclerosis, thrombosis, and vascular biology* 41.4 (2021), pp. 1446–1458.

[122] Quoc Le and Tomas Mikolov. "Distributed representations of sentences and documents". In: *International conference on machine learning*. PMLR. 2014, pp. 1188–1196.

[123] Isabelle Leang et al. "Dynamic task weighting methods for multi-task networks in autonomous driving systems". In: *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2020, pp. 1–8.

[124] Benjamin Yee Shing Li, Lam Fat Yeung, and Genke Yang. "Pathogen host interaction prediction via matrix factorization". In: *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2014, pp. 357–362.

[125] Jie Li et al. "Network-based identification of microRNAs as potential pharmacogenomic biomarkers for anticancer drugs". In: *Oncotarget* 7.29 (2016), pp. 45584–45596.

[126] Jin Li et al. "Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction". In: *Bioinformatics* 36.8 (2020), pp. 2538–2546.

[127] Lei Li et al. "SCMFMDA: Predicting microRNA-disease associations based on similarity constrained matrix factorization". In: *PLoS computational biology* 17.7 (2021), e1009165.

[128] Wenhui Li et al. "Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus". In: *Nature* 426.6965 (2003), pp. 450–454.

[129] Yang Li et al. "HMDD v2. 0: a database for experimentally supported human microRNA and disease associations". In: *Nucleic acids research* 42.D1 (2014), pp. D1070–D1074.

[130] Yiwei Li. "Computational Methods for Predicting Protein-protein Interactions and Binding Sites". In: (2020).

[131] Yiwei Li and Lucian Ilie. "Predicting protein–protein interactions using sprint". In: *Protein-Protein Interaction Networks 2020*. Springer, pp. 1–11.

[132] Zhong Li et al. "GCSENet: A GCN, CNN and SENet ensemble model for microRNA-disease association prediction". In: *PLOS Computational Biology* 17.6 (2021), e1009048.

[133] Yi Lin et al. "Characterization of microRNA expression profiles and the discovery of novel microRNAs involved in cancer during human embryonic development". In: *PloS one* 8.8 (2013), e69230.

[134] Dan Liu et al. "Predicting virus-host association by Kernelized logistic matrix factorization and similarity network fusion". In: *BMC bioinformatics* 20.16 (2019), pp. 1–10.

[135] Na Liu et al. "The evolution and functional diversification of animal microRNA genes". In: *Cell research* 18.10 (2008), pp. 985–996.

[136] Wang Liu-Wei et al. "DeepViral: prediction of novel virus–host interactions from protein sequences and infectious disease phenotypes". In: *Bioinformatics* (Mar. 2021). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab147.

[137] Cristian D Loaiza and Rakesh Kaundal. "PredHPI: an integrated web server platform for the detection and visualization of host–pathogen interactions using sequence-based methods". In: *Bioinformatics* (2020).

[138] Krystal K Lum and Ileana M Cristea. "Proteomic approaches to uncovering virus–host protein interactions during the progression of viral infection". In: *Expert review of proteomics* 13.3 (2016), pp. 325–340.

[139] Sali Lv et al. "A novel method to quantify gene set functional association based on gene ontology". In: *Journal of The Royal Society Interface* 9.70 (2012), pp. 1063–1072.

[140] Yingli Lv et al. "Identifying novel associations between small molecules and miRNAs based on integrated molecular networks". In: *Bioinformatics* 31.22 (2015), pp. 3638–3644.

[141] Yingjun Ma, Tingting He, Yu-Ting Tan, et al. "Seq-BEL: Sequence-based Ensemble Learning for Predicting Virus-human Protein-protein Interaction". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2020).

[142] Zheng Ma, Xiugang Wu, and Han Xu. "Understanding the degree distribution pattern of protein-protein interaction networks". In: *2015 International Symposium on Bioelectronics and Bioinformatics (ISBB)*. IEEE. 2015, pp. 43–46.

[143] Neel S Madhukar et al. "A Bayesian machine learning approach for drug target identification using diverse data types". In: *Nature communications* 10.1 (2019), pp. 1–14.

[144] Nathan Mantel. "Evaluation of survival data and two new rank order statistics arising in its consideration". In: *Cancer Chemother Rep* 50 (1966), pp. 163–170.

[145] Shawn Martin, Diana Roe, and Jean-Loup Faulon. "Predicting protein–protein interactions using signature products". In: *Bioinformatics* 21.2 (2005), pp. 218–226.

[146] L Martin-Gutierrez et al. "Two shared immune cell signatures stratify patients with Sjögren's syndrome and systemic lupus erythematosus with potential therapeutic implications." In: *Arthritis & Rheumatology (Hoboken, NJ)* (2021).

[147] John S Mattick and Igor V Makunin. "Small regulatory RNAs in mammals". In: *Human molecular genetics* 14.suppl_1 (2005), R121–R132.

[148] Suyu Mei. "Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins". In: *PLoS One* 8.11 (2013), e79606.

[149] Suyu Mei and Kun Zhang. "In silico unravelling pathogen-host signaling cross-talks via pathogen mimicry and human protein-protein interaction networks". In: *Computational and structural biotechnology journal* 18 (2020), pp. 100–113.

[150] Suyu Mei and Hao Zhu. "A novel one-class SVM based negative data sampling method for reconstructing proteome-wide HTLV-human protein interaction networks". In: *Scientific reports* 5.1 (2015), pp. 1–13.

[151] Xiaolan Mo et al. "Early prediction of clinical response to etanercept treatment in juvenile idiopathic arthritis using machine learning". In: *Frontiers in pharmacology* 11 (2020), p. 1164.

[152] Pieter Moris. *Python script and package for Gene Ontology enrichment analysis*. https://pmoris.github.io/goscripts/_build/html/source/README.html.

[153] Søren Mørk et al. "Protein-driven inference of miRNA–disease associations". In: *Bioinformatics* 30.3 (2014), pp. 392–397.

[154] *Multi-layer Perceptron classifier*. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.

[155] Sara G Murray et al. "Automated and flexible identification of complex disease: building a model for systemic lupus erythematosus using noisy labeling". In: *Journal of the American Medical Informatics Association* 26.1 (2019), pp. 61–65.

[156] Victor Neduva and Robert B Russell. "Peptides mediating interaction networks: new leads at last". In: *Current opinion in biotechnology* 17.5 (2006), pp. 465–471.

[157] Saul B Needleman and Christian D Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *Journal of molecular biology* 48.3 (1970), pp. 443–453.

[158] Esmaeil Nourani, Farshad Khunjush, and Saliha Durmuş. "Computational prediction of virus–human protein–protein interactions using embedding kernelized heterogeneous data". In: *Molecular BioSystems* 12.6 (2016), pp. 1976–1986.

[159] Ilia Nouretdinov et al. "Determining confidence of predicted interactions between HIV-1 and human proteins using conformal method". In: *Biocomputing 2012*. World Scientific, pp. 311–322.

[160] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453.

[161] Randal S. Olson. *ReliefF 0.1.2*. https://pypi.org/project/ReliefF/. Released: Mar 20, 2016.

[162] Xiaoyong Pan and Hong-Bin Shen. "Scoring disease-microRNA associations by integrating disease hierarchy into graph convolutional networks". In: *Pattern Recognition* 105 (2020), p. 107385.

[163] Sandra Gofinet Pasoto, Victor Adriano de Oliveira Martins, and Eloisa Bonfa. "Sjögren's syndrome and systemic lupus erythematosus: links and risks". In: *Open Access Rheumatology: Research and Reviews* 11 (2019), p. 33.

[164] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019), pp. 8026–8037.

[165] Yasset Perez-Riverol et al. "Making proteomics data accessible and reusable: current state of proteomics databases and repositories". In: *Proteomics* 15.5-6 (2015), pp. 930–950.

[166] Eskild Petersen et al. "Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics". In: *The Lancet Infectious Diseases* 20.9 (2020), e238–2244.

[167] Richard Peto and Julian Peto. "Asymptotically efficient rank invariant test procedures". In: *Journal of the Royal Statistical Society: Series A (General)* 135.2 (1972), pp. 185–207.

[168] Janet Piñero et al. "The DisGeNET knowledge platform for disease genomics: 2019 update". In: *Nucleic acids research* 48.D1 (2020), pp. D845–D855.

[169] Sune Pletscher-Frankild et al. "DISEASES: Text mining and data integration of disease–gene associations". In: *Methods* 74 (2015), pp. 83–89.

[170] Jia Qu et al. "In Silico prediction of small molecule-miRNA associations based on the HeteSim algorithm". In: *Molecular Therapy-Nucleic Acids* 14 (2019), pp. 274–286.

[171] Jia Qu et al. "Inferring potential small molecule–miRNA association based on triple layer heterogeneous network". In: *Journal of cheminformatics* 10.1 (2018), p. 30.

[172] Sameer Quazi et al. "Artificial intelligence and machine learning in medicinal chemistry and validation of emerging drug targets". In: *Advancements in Controlled Drug Delivery Systems* (2022), pp. 27–43.

[173] Jennifer Raisch, Arlette Darfeuille-Michaud, and Hang Thi Thu Nguyen. "Role of microRNAs in the immune system, inflammation and cancer". In: *World journal of gastroenterology: WJG* 19.20 (2013), pp. 2985–2996.

[174] Dhanesh Ramachandram and Graham W Taylor. "Deep multimodal learning: A survey on recent advances and trends". In: *IEEE signal processing magazine* 34.6 (2017), pp. 96–108.

[175] Rodrigo D Requião et al. "Viruses with different genome types adopt a similar strategy to pack nucleic acids based on positively charged protein domains". In: *Scientific reports* 10.1 (2020), pp. 1–12.

[176] Guillermo Rodrigo, José-Antonio Daròs, and Santiago F Elena. "Virus-host interactome: putting the accent on how it changes". In: *Journal of proteomics* 156 (2017), pp. 1–4.

[177] Juan José Rodriguez, Ludmila I Kuncheva, and Carlos J Alonso. "Rotation forest: A new classifier ensemble method". In: *IEEE transactions on pattern analysis and machine intelligence* 28.10 (2006), pp. 1619–1630.

[178] Nancy J Roizen and David Patterson. "Down's syndrome". In: *The Lancet* 361.9365 (2003), pp. 1281–1289.

[179] Maurice WJ de Ronde et al. "Study design and qPCR data analysis guidelines for reliable circulating miRNA biomarker experiments: a review". In: *Clinical chemistry* 64.9 (2018), pp. 1308–1318.

[180] Johan Rung and Alvis Brazma. "Reuse of public genome-wide gene expression data". In: *Nature Reviews Genetics* 14.2 (2013), pp. 89–99.

[181] Sepideh Sadegh et al. "Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing". In: *Nature communications* 11.1 (2020), pp. 1–9.

[182] Debashis Sahoo et al. "Extracting binary signals from microarray time-course data". In: *Nucleic acids research* 35.11 (2007), pp. 3705–3712.

[183] Harpreet Kaur Saini, Sam Griffiths-Jones, and Anton James Enright. "Genomic analysis of human microRNA transcripts". In: *Proceedings of the National Academy of Sciences* 104.45 (2007), pp. 17719–17724.

[184] Kioomars Saliminejad et al. "An overview of microRNAs: Biology, functions, therapeutics, and analysis methods". In: *Journal of Cellular Physiology* 234 (5 May 2019), pp. 5451–5465. ISSN: 10974652. DOI: 10.1002/jcp.27486.

[185] Alessandro Salvi et al. "Analysis of a nanoparticle-enriched fraction of plasma reveals miRNA candidates for Down syndrome pathogenesis". In: *International journal of molecular medicine* 43.6 (2019), pp. 2303–2318.

[186] Steven L Salzberg. "On comparing classifiers: Pitfalls to avoid and a recommended approach". In: *Data mining and knowledge discovery* 1.3 (1997), pp. 317–328.

[187] Steven L Salzberg. "Open questions: How many genes do we have?" In: *BMC biology* 16.1 (2018), pp. 1–3.

[188] Debasree Sarkar and Sudipto Saha. "Machine-learning techniques for the prediction of protein–protein interactions". In: *Journal of biosciences* 44.4 (2019), pp. 1–12.

[189] R Schickel et al. "MicroRNAs: key players in the immune system, differentiation, tumorigenesis and cell death". In: *Oncogene* 27.45 (2008), pp. 5959–5974.

[190] Lynn Marie Schriml et al. "Disease Ontology: a backbone for disease semantic integration". In: *Nucleic acids research* 40.D1 (2012), pp. D940–D946.

[191] Jessica Schulz et al. "Meta-analyses identify differentially expressed microRNAs in Parkinson's disease". In: *Annals of neurology* 85.6 (2019), pp. 835–851.

[192] *Scikit-learn Average Precision score*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html.

[193] *Scikit-learn Random Forest classifier*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[194] Jian Shang et al. "Cell entry mechanisms of SARS-CoV-2". In: *Proceedings of the National Academy of Sciences* 117.21 (2020), pp. 11727–11734.

[195] Cong Shen et al. "Identification of small molecule–miRNA associations with graph regularization techniques in heterogeneous networks". In: *Journal of Chemical Information and Modeling* 60.12 (2020), pp. 6709–6721.

[196] Alicia E Smith and Ari Helenius. "How viruses enter animal cells". In: *Science* 304.5668 (2004), pp. 237–242.

[197] Gregory A Smith and Lynn W Enquist. "Break ins and break outs: viral interactions with the cytoskeleton of Mammalian cells". In: *Annual Review of cell and developmental biology* 18 (2002), pp. 135–61.

[198] Christina F Spiropoulou et al. "New World arenavirus clade C, but not clade A and B viruses, utilizes α-dystroglycan as its major receptor". In: *Journal of virology* 76.10 (2002), pp. 5140–5146.

[199] Justin Stebbing et al. "Mechanism of baricitinib supports artificial intelligence-predicted testing in COVID-19 patients". In: *EMBO molecular medicine* 12.8 (2020), e12697.

[200] Padhmanand Sudhakar, Kathleen Machiels, and Severine Vermeire. "Computational Biology and Machine Learning Approaches to Study Mechanistic Microbiomehost Interactions". In: (2020).

[201] Tanlin Sun et al. "Sequence-based prediction of protein protein interaction using a deep-learning algorithm". In: *BMC bioinformatics* 18.1 (2017), pp. 1–8.

[202] *SVM classifier*. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

[203] Damian Szklarczyk et al. "STRING v10: protein–protein interaction networks, integrated over the tree of life". In: *Nucleic acids research* 43.D1 (2015), pp. D447–D452.

[204] Damian Szklarczyk et al. "The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets". In: *Nucleic acids research* 49.D1 (2021), pp. D605–D612.

[205] Xinru Tang et al. "Multi-view multichannel attention graph convolutional network for miRNA–disease association prediction". In: *Briefings in Bioinformatics* 22.6 (2021), bbab174.

[206] Ryan J Urbanowicz et al. "Relief-based feature selection: Introduction and review". In: *Journal of biomedical informatics* 85 (2018), pp. 189–203.

[207] Wataru Usuba et al. "Circulating miRNA panels for specific and early detection in bladder cancer". In: *Cancer science* 110.1 (2019), pp. 408–419.

[208] Erika Van Nieuwenhove et al. "Machine learning identifies an immunological pattern associated with multiple juvenile idiopathic arthritis subtypes". In: *Annals of the Rheumatic Diseases* 78.5 (2019), pp. 617–628.

[209] Naveen Vankadari and Jacqueline A Wilce. "Emerging COVID-19 coronavirus: glycan shield and structure prediction of spike glycoprotein and its interaction with human CD26". In: *Emerging microbes & infections* 9.1 (2020), pp. 601–604.

[210] Marc Vaudel et al. "Exploring the potential of public proteomics data". In: *Proteomics* 16.2 (2016), pp. 214–225.

[211] Dan Ventura and Sean Warnick. "A theoretical foundation for inductive transfer". In: *Brigham Young University, College of Physical and Mathematical Sciences* 19 (2007).

[212] Akbar K Waljee et al. "Development and validation of machine learning models in prediction of remission in patients with moderate to severe Crohn disease". In: *JAMA network open* 2.5 (2019), e193721–e193721.

[213] Derek Walsh and Ian Mohr. "Viral subversion of the host protein synthesis machinery". In: *Nature Reviews Microbiology* 9.12 (2011), pp. 860–875.

[214] Chun-Chun Wang and Xing Chen. "A unified framework for the prediction of small molecule–MicroRNA association based on cross-layer dependency inference on multilayered networks". In: *Journal of chemical information and modeling* 59.12 (2019), pp. 5281–5293.

[215] Chun-Chun Wang, Chi-Chi Zhu, and Xing Chen. "Ensemble of kernel ridge regression-based small molecule–miRNA association prediction in human disease". In: *Briefings in Bioinformatics* 23.1 (2022), bbab431.

[216] Chun-Chun Wang et al. "Prediction of potential miRNA–disease associations based on stacked autoencoder". en. In: *Briefings in Bioinformatics* 23.2 (Mar. 2022), bbac021. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbac021. URL: https://academic.oup.com/bib/article/doi/10.1093/bib/bbac021/6529883 (visited on 06/10/2022).

[217] Chun-Chun Wang et al. "RFSMMA: a new computational model to identify and prioritize potential small molecule–mirna associations". In: *Journal of chemical information and modeling* 59.4 (2019), pp. 1668–1679.

[218] Daixin Wang, Peng Cui, and Wenwu Zhu. "Structural deep network embedding". In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 2016, pp. 1225–1234.

[219] Dong Wang et al. "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases". In: *Bioinformatics* 26.13 (2010), pp. 1644–1650.

[220] Lei Wang et al. "LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities". In: *PLoS computational biology* 15.3 (2019), e1006865.

[221] Nianshuang Wang et al. "Structure of MERS-CoV spike receptor-binding domain complexed with human receptor DPP4". In: *Cell research* 23.8 (2013), pp. 986–993.

[222] Shu-Hao Wang et al. "Dual-Network Collaborative Matrix Factorization for predicting small molecule-miRNA associations". In: *Briefings in Bioinformatics* 23.1 (2022), bbab500.

[223] Weili Wang et al. "A network-based integrated framework for predicting virus–prokaryote interactions". In: *NAR genomics and bioinformatics* 2.2 (2020), lqaa044.

[224] Yixin Wang et al. "Hemicastration induced spermatogenesis-related DNA methylation and gene expression changes in mice testis". In: *Asian-Australasian journal of animal sciences* 31.2 (2018), p. 189.

[225] Bernard L Welch. "The generalization of 'STUDENT'S'problem when several different population varlances are involved". In: *Biometrika* 34.1-2 (1947), pp. 28–35.

[226] Fabian Wendt, Emanuela S Milani, and Bernd Wollscheid. "Elucidation of host-virus surfaceome interactions using spatial proteotyping." In: *Advances in Virus Research* 109 (2021), pp. 105–134.

[227] Emma S Winkler et al. "SARS-CoV-2 infection of human ACE2-transgenic mice causes severe lung inflammation and impaired function". In: *Nature immunology* 21.11 (2020), pp. 1327–1335.

[228] David S Wishart et al. "DrugBank 5.0: a major update to the DrugBank database for 2018". In: *Nucleic acids research* 46.D1 (2018), pp. D1074–D1082.

[229] Guanming Wu, Xin Feng, and Lincoln Stein. "A human functional protein interaction network and its application to cancer data analysis". In: *Genome biology* 11.5 (2010), R53.

[230] Qiu Xiao et al. "A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations". In: *Bioinformatics* 34.2 (2018), pp. 239–248.

[231] Boya Xie et al. "miRCancer: a microRNA–cancer association database constructed by text mining on literature". In: *Bioinformatics* 29.5 (2013), pp. 638–644.

[232] Xintian Xu et al. "Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission". In: *Science China Life Sciences* 63.3 (2020), pp. 457–460.

[233] Ping Xuan et al. "Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors". In: *PloS one* 8.8 (2013), e70204.

[234] Xiaodi Yang et al. "Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method". In: *Computational and structural biotechnology journal* 18 (2020), pp. 153–161.

[235] Yang Yang, Ryan N Lichtenwalter, and Nitesh V Chawla. "Evaluating link prediction methods". In: *Knowledge and Information Systems* 45.3 (2015), pp. 751–782.

[236] Zhen Yang et al. "dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers". In: *Nucleic acids research* 45.D1 (2017), pp. D812–D818.

[237] Curtis L Yeager et al. "Human aminopeptidase N is a receptor for human coronavirus 229E". In: *Nature* 357.6377 (1992), pp. 420–422.

[238] Francisco José Zapatero-Belinchón, Belén Carriquí-Madroñal, and Gisa Gerold. "Proximity labeling approaches to study protein complexes during virus infection." In: *Advances in Virus Research* 109 (2021), pp. 63–104.

[239] Xiangxiang Zeng et al. "Prediction of potential disease-associated microRNAs by using neural networks". In: *Molecular Therapy-Nucleic Acids* 16 (2019), pp. 566–575.

[240] Xiangxiang Zeng et al. "Target identification among known drugs by deep learning from heterogeneous networks". In: *Chemical Science* 11.7 (2020), pp. 1775–1797.

[241] Qianqian Zhang et al. "Molecular mechanism of interaction between SARS-CoV-2 and host cells and interventional therapy". In: *Signal Transduction and Targeted Therapy* 6.1 (2021), pp. 1–19.

[242] Wenyong Zhang, James E Dahlberg, and Wayne Tam. "MicroRNAs in tumorigenesis: a primer". In: *The American journal of pathology* 171.3 (2007), pp. 728–738.

[243] Yi Zhang et al. "MSFSP: a novel miRNA–disease association prediction model by federating multiple-similarities fusion and space projection". In: *Frontiers in Genetics* 11 (2020), p. 389.

[244] Kai Zheng et al. "DBMDA: A Unified Embedding for Sequence-Based miRNA Similarity Measure with Applications to Predict and Validate miRNA-Disease Associations". In: *Molecular Therapy-Nucleic Acids* 19 (2020), pp. 602–611.

[245] Lu Zheng and Young Chun Ko. "Application of big data adaptive semi-supervised clustering method based on deep learning". In: *Journal of Computational Methods in Sciences and Engineering* Preprint (2022), pp. 1–15.

[246] Xiang Zhou et al. "A generalized approach to predicting protein-protein interactions between virus and host". In: *BMC genomics* 19.6 (2018), pp. 69–77.

[247] Fuzhen Zhuang et al. "A comprehensive survey on transfer learning". In: *Proceedings of the IEEE* 109.1 (2020), pp. 43–76.

Ngan Dong

Curriculum Vitae

PERSONAL DETAILS

Birth	July 24, 1988
Address	Hanoi, Vietnam
Phone	+(84) 915-707-585
Mail	dong@l3s.de
Website	http://l3s.de/~dong/
G. Scholar	https://scholar.google.com/citations?user=M4zbOcsAAAAJ

EDUCATION

PhD Computer Science 2019 - 2022

Leibniz University Hannover

Research Topic: "Deep learning for precision medicine"

MSc. Computer Science 2011-2013

Washington State University

Thesis: "Natural Language Generation from Graphs"

BSc. in Computer Science, Honor program 2006-2010

VNU University of Engineering and Technology

EXPERTISE

- More than 5 years of developing Machine Learning algorithms at industry scale for Recommendation systems and various Natural Language Processing tasks.
- Over 3 years of working on Machine Learning models for Biological problems which focus on jointly learning from multiple sources, Graph Representation Learning techniques, Multi-task learning, and Feature Selection.

WORK EXPERIENCE

Research Assistant 2019-Present

Leibniz University Hannover, Full-time

- Working on the PRESENt project[1] which aims at integrating clinical, biological, and big data research to advance our understanding of norovirus gastroenteritis.
- Performed survival analysis, differential analysis, feature selection on RNA-sequencing and protein expression data.
- Developed state-of-the-art models for the protein-protein interaction prediction, miNRA-disease association prediction problems that focus on learning from multiple data sources, network analysis, multitask learning, graph representation learning, and feature selection.
- Developed a cochlear implant outcome prediction model from real patient data. The work involved heavy data preprocessing tasks, feature extraction, feature selection, and model development.

[1] http://www.translationsallianz.de/train-plattformen/train-projects/present/

- Perform some others data retrieval, data pre-processing tasks and specialized analysis on the results retrieved from wet-lab experiments given by our biologist partners.

Research Engineer(Machine Learning) 2014-2019

FPT Technology Research Institute, Full-time

- Being one of the first ML researchers working on Natural Language Understanding module for FPT.AI[2] - one of the first and largest comprehensive AI platforms in Vietnam.
- Worked on (i) Recommendation systems for e-commerce websites and online news, (ii) Sentiment classification for electronic retailer, (iii) User segmentation, and (iv) online-news topic modeling

Software Developer 2013-2014

Citigo Joint stock Company., Full-time

Work as a full-stack developer on out-sourcing projects for Australian customers. Responsible for maintaining the existing systems, developing, testing and deploying new functionality on customers' production servers.

Research Assistant 2011-2013

Artificial Intelligence laboratory, WSU

- Develop GNLG - a Natural Language Generation (NLG) system for Resource Description Framework (RDF) Graphs.
- Build ontology models for cooking recipes and data from a defense-related project about people in a military- controlled area

Teaching Assistant 2010-2011

Computer Science Department, VNU University of Engineering and Technology

- Teaching Assistant for various bachelor level programming courses.

Research Assistant 2009-2010

Computer Science Department, VNU University of Engineering and Technology

- Research on POS taggers for Vietnamese using Conditional Random Fields and Hidden Markov Models.
- Develop a Vietnamese collocation extractor using different statistical methods

SKILLS

Languages	Vietnamese (mother tongue), English (fluent), German (basic)
Code related	Python, PyTorch, pytorch-geometric, sklearn, pandas, numpy, networkx

PUBLICATIONS

Ngan Dong, Johanna Schrader, Stefanie Mücke, Megha Khosla, "*A Message Passing framework with Multiple data integration for miRNA-Disease association prediction*", to be published at Scientific Reports, 2022.

Ngan Dong, Stefanie Mücke, Megha Khosla, "*MuCoMiD: A Multitask graph Convolutional Learning Framework for miRNA-Disease Association Prediction*", in IEEE/ACM Transactions on Computational Biology and Bioinformatics, doi: 10.1109/TCBB.2022.3176456, 2022.

Ngan Dong, Graham Brogden, Gisa Gerold, Megha Khosla, "*A multitask transfer learning framework for the prediction of virus-human protein-protein interactions*", BMC Bioinformatics, 2021.

Ngan Dong, Megha Khosla, "*A Multitask Convolutional Learning Framework for miRNA-Disease Association Prediction*", BIOKDD 2021.

[2]https://fpt.ai/

Ngan Dong, Megha Khosla,, "*A multitask transfer learning framework for Novel virus-human protein interactions*", ICLR Workshop on AI for Public Health, 2021.

Ngan Dong, Megha Khosla, "*Towards a consistent evaluation of miRNA-disease association prediction models.*", 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2020.

Ngan Dong, Megha Khosla. "*Revisiting Feature Selection with Data Complexity for Biomedicine.*" 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE).

Kim Anh Nguyen, **Ngan Dong**, Cam Tu Nguyen "*Attentive Neural Network for Named Entity Recognition in Vietnamese*", 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF). IEEE, 2019.

Ngan Dong, Larry Holder, "*Natural Language Generation from Graphs*", International Journal of Semantic Computing.Vol. 8, No. 3, pp. 335-384, 2014

ACADEMIC SERVICES

Teaching assistant for Machine Learning for Graphs course (2021)
Seminar supervisor for Artificial Intelligence course (2021,2022)
Reviewed papers at CHIL 2022, BMC Bioinformatics
Msc. students supervised/co-supervised
Johanna Schrader - Thesis title: Application of Graph Structure Learning in Biological Data Analysis (2021)
Luo Yi - Thesis title: At the Interface between Biomedical Research and Software engineering (2021)

HONOR AND AWARDS

August 2011 – May 2013: Vietnam Education Foundation (VEF) scholarship, which was funded by the US government.

July 2010: Scholarship for Digital Signal Processing Academy Summer School

Dec 2009: Outstanding female student in Information Technology
Scholarship from the Ministry of Information and Communication, Vietnam

Oct 2007: VNU University of Engineering and Technology Outstanding student

REFERENCES

Available upon request.