# Prediction of Missing and Spurious Links in Biological Networks

Name: Alex Khawalid
Studentnumber: 10634207
Supervisor: Putri van der Linden

January 8, 2021

**Abstract**

Elsevier is working in close collaboration with the Rochester General Hos- pital in harnessing prior knowledge from medical literature to assemble and analyse biological networks for the purpose of identifying intervention targets in complex chronic illnesses. Although this prior knowledge of regulatory interac- tions is being extracted from Elseviers sizable corpus of over 5 million journal papers, the complexity of biological signalling is such much remains to be dis- covered. Hypothesis generation informed on the basis of regulatory interactions that are currently reported is a critical component in the optimal planning of new experiments. The prediction of spurious or indirect associations (false pos- itives), as well as the absence of functionally important direct interactions (false negatives) has been proposed based on adherence to known fundamental topo- logical properties of biological networks [1], albeit in undirected networks. The aim of this project is to extend this work to the prediction of novel interactions based on directed regulatory interactions and their kineticproperties. Applying Alon [2] regulatory motifs as basic building blocks is one potential approach. Machine learning techniques would be applied to artificial networks with known defects to create predictive models for inferring the such novel interactions. The results of the project will be summarized in a paper that will be submitted in high-impact peer review journal or conference.

## 1 Introduction

Network data is not always reliable, especially for biological networks. Given a knowledge graph, and experimental data from a biological system, how can we assign a probability to each edge? One approach relies on traditional network analysis methods such as Matrix Factorization to describe a graph through mathematical analysis, encompassing higher order features as well as local features. However, deep learning has shown promise in the field of bioinformatics through multiple endeavors. A recent example is AlphaFold, a deep learning

system that is able to predict protein folding. More relevantly, graph neural networks and graph embeddings have shown competitive performance on link prediction tasks. Furthermore, generative models have also been shown to be effective at probablistic link prediction tasks. One of the reasons is that generative models such as VAEs tend to have better interpolation on data density estimations (which are an essential part of link prediction). This report will explore applications of deep learning methods for probablistic link prediction.

# 2 Methods

## 2.1 Matrix Factorization

Mathematical analysis (e.g. GraRep).

## 2.2 Network embeddings

The algorithm ignores directionality of data. Edge embeddings are created by using an operator (e.g. Hadamard operator) on two node embeddings. Use of these edge embeddings is optional.

Node2vec edge embeddings seem to be a bad representation of edges. Using the edge embeddings gives us an F1 score of 0. However, when concatenating node embeddings and using these as features, this gives us an F1 score of 0.78. Similarly so for Struc2Vec.

## 2.3 Graph Neural Networks

## 2.4 Convolutional Graph Neural Networks

Based on the graph neural networks survey, I don't think spectral ConvGNNs are viable, spectral based methods are only applicable to undirected graphs[**?**, p. 11]. Furthermore, they're less efficient than spatial ConvGNNs, so from the overview of methods i'll give priority to non-spectral methods.

## 2.5 Generative Graph Neural Networks

# 3 Experimental Setup

## 3.1 Data

The datasets used in this experiment concern protein concentrations and their influences on each other. The models are initially tested with 50-100 nodes and should scale to around 5000 nodes.

Human tissue network data was used for initial testing, from the paper (Predicting multicellular function throughmulti-layer tissue networksMarinka), the tissue.edges.

### 3.2 Evaluation

The evaluation method used is the F-score, this is especially important as link prediction is essentially a binary classification problem (edges exist or edges do not exist) with a huge class imbalance.

## 4 Results

### 4.1 Node2vec and Struc2vec

Both perform similarly in terms of F-score.