Microbe-Drug Association Prediction using Graph Neural Networks

Anushka Sivakumar[1*], J. Angel Arul Jothi[2]

1 Department of Computer Science, Birla Institute of Technology and Science Pilani Dubai Campus, Dubai, UAE
f20190208@dubai.bits-pilani.ac.in

2 Department of Computer Science, Birla Institute of Technology and Science Pilani Dubai Campus, Dubai, UAE
angeljothi@dubai.bits-pilani.ac.in

Abstract: Human microorganisms play a crucial factor in regulating the immune system, producing hormones, providing energy, participating in human metabolism, and carrying out important bodily functions using a diverse range of chemical reactions, some of which cannot be carried out by human enzymes. While our system maintains a healthy relationship with good microbes, several disease-related states have been linked to the microbes. Identifying microbe-drug associations is useful in gaining insights and understanding the mechanism of their relationship. The information can be used for developing and repurposing drugs, clinical treatments and personalized medicine. In this work, various biological datasets are leveraged to construct heterogenous networks for microbe-drug known associations and microbe-disease-disease-drug transitive associations, as well as feature matrices containing information on microbe functional similarity and microbe genome similarity, and drug structure similarity and drug side-effect-based similarity. The input data is then fed to Graph Neural Network models (GAT & GCN) that creates an encoded embedding which is put through a deep neural network classifier to accurately predict potential microbe-drug associations. The proposed model is evaluated to compare GCN and GAT and quantify the overall suitability of such a model to predict scores for microbe-drug pairs.

Keywords: Graph Neural Networks, Metagenomics, Microbe-drug associations, Prediction model

AMS Subject classification:

* *Corresponding Author*

Contents

1. Introduction

Microbe or microorganism is a microscopic living organism that can be categorized as a single-cell or multi-cell organism [1]. The human microbiome represents a complex community of trillions of microorganisms made up of bacteria, archea, protozoa, virus, fungi, etc. [2]. These microbial communities are found in organs such as the skin, gastrointestinal tract, lung, oral cavity, vagina, and other tissues [1]. While a handful of human microbes have beneficial functions and play a fundamental role in the human environment such as homeostasis, improvement of metabolism and immunity, and synthesis of essential vitamins, some microbes are also the cause of multiple diseases such as Inflammatory Bowel Disease (IBD), Severe Acute Respiratory Syndrome (SARS), etc. that directly affect human health [2, 3]. With rapid mutations and a rise in resistance, there is a disruption in the study relationship between the microbiome and body cells, which can be linked to several diseases [4].

The importance of microbe-drug predictions and understanding their complex interactions is seen in various settings such as personalized medicine, microbe-derived therapy, clinical treatment, precision medicine, and drug discovery. Further, uncovering potential associations can aid in the research on the affects of microbial metabolism on response to drugs, and research on drug repositioning and combinations to help deal with the rising antibiotic resistance [1, 2, 5].

There are numerous drawbacks with the conventional methods of identifying possible associations between drugs and microbes with the prominent ones being the amount of time, labour, and cost of experimenting and developing the calculation models [5]. Additionally, the difficulty in selecting target microbes further slows the progress of developing new drugs [6].

Recent advancements in domains of machine learning and deep learning have made it possible to accurately predict microbe-drug associations and complement traditional wet-lab experiments while accounting for the aforementioned shortcomings [7]. Additionally, the increasing availability of biological data such as the Microbe Drug Association Database (MDAD) [8] containing a large number of experimentally validated associations between microbes and drugs, Human Microbe Disease Association Database (HMDAD) [9] containing microbe-disease associations and BIOSNAP DCh-Miner dataset [10] containing drug-disease associations, allows for various calculation methods to be proposed to identify possible latent and transitive links between microbes and drugs.

A lot of underlying relationships between data can be modeled as graphs and among various approaches, graph or network embedded representation learning seems to be a popular and effective approach to process graphs due to its ability to accurately capture structural information of the network [1, 11]. Graphs are considered to be complex data without fixed strcuture or ordering. Graph Neural Networks (GNN) are supervised deep learning models that are applied to various graph domains [11]. Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT) are

a type of GNN that show great potential in capturing and embedding topological information and modeling complex graph data. GNNs are popularly used for various graph analysis tasks such as node classification, link prediction, and clustering [12].

All things considered, the proposed framework puts forward the following novel contributions by combining the powerful encoded embedding capabilities of GNNs.

1. Construction of a hetergenous network H of size $(nd + nm \times nd + nm)$ where nm denotes the number of microbes and nm denotes the number of drugs. H is constructed based on the adjacency matrix A of size $(nm \times nd)$ which forms a bipartite network, of known associations between microbes and drugs and transitive associations between the set of microbes-disease and disease-drugs links. Additionally, two homogenous networks are constructed, one for microbe-microbe interactions Pm of size $(nm \times nm)$, and one for drug-drug interactions Pd of size $(nd \times nd)$.
2. A feature matrix F is constructed of size $(nd+nm \times 2*(nd+nm))$. It consists of multiple attributes such as the microbe functional similarity, microbe genome sequence feature, linear mean of Gaussian Interaction Profile similarity, side-effect based similarity and structural similarity for drugs, and random-walk-restart features for drugs.
3. A GAT and GCN network are compared in their ability to encode embeddings and accurately predict possible associations between microbes and drugs by predicting scores between the pairs.The heteregenous network H and the features F are fed as input to the model to uncover latent graph predictions. The Deep Neural Network (DNN) based classifier is used for the microbe-drug association predictions.

2. Related Works

Numerous studies have been conducted on effective and efficient microbe-drug association prediction techniques. This review aims to analyze the various methods implemented on different datasets and identify the advantages and disadvantages of each methodology which help to guide this work.

The various methods surveyed begin by constructing bipartite and heterogenous networks to map the associations between microbes and drugs. A bipartite network maps association from a set of microbes to a set of drugs (undirected). A hetergoneous network is a bipartite network along with links between microbes and links between drugs in the graph (i.e homogenous networks). Microbe-microbe interactions are typically mapped based on Gaussian Interaction Profile (GIP), random walk restart, microbe functional similarity and or microbe genome similarity. Drug-drug interactions are commonly mapped based on GIP, random walk restart, and or drug structure similarity.

Long et al. proposed an Ensemble Graph Attention Network Microbe-Drug Association (EGATMDA) model. Three input graphs were generated representing microbe-drug bipartite network, heterogenous network and a microbe-disease-drug heterogenous network. Each of the graphs were fed as input to the Graph Convolution Network (GCN) along with input feature matrices. The microbe feature matrix represented microbe genome similarity and the drug feature matrix integrated drug structure similarity and drug GIP kernel similarity. Combining the output of the GCN with node attention calculations, the GAT layer captured the graph level attention. Finally, the three GNNs were aggregated and decoded to form the final microbe-drug association score matrix. [6]

By accounting for multiple node features to improve model performance, Tan et al. proposed Graph attention and Sparse Auto-encoder (SAE) microbe drug association GSAMDA. An heterogenous network with GIP kernel similarity and Hamming Interaction Profile similarity calculated for microbes, and drugs was generated. The GAT encoder-decoder model learns the topological representations and the output matrix is then fed to the SAE along with microbe and drug feature matrices where the latter additionally made use of disease cosine similarity. The SAE learns attribute representations and intrinsic features. The output matrices from the GAT and SAE are combined to generate the final association scores matrix. [13]

A simple architecture based on GCN with a Conditional Random Field (CRF) was proposed by Long et al. to capture the links between microbes and drugs. By applying Random walk restart on these networks, individual feature matrices are generated. The heterogenous network and feature matrices are inputted to the GCN based encoder which has a CRF layer to aggregate representations of nodes and assign attention scores. After passing through the decoder, the final reconstructed microbe-drug association bipartite network is achieved. [7]

Contrary to the popular GNN models, Zhu et al. proposed a model based on Laplachian Regularized Least Squares (LRLS) to identify associations. Diagonal matrices were generated based on similarity matrices. Laplachian operation on the normalized matrices after which a minimization cost function based on LRLS algorithm generates two prediction matrices the linear mean of which gives the final microbe-drug associations. [2]

Long and Luo proposed a model applying association mining to predict microbe-drug associations. The heterogeneous network was mapped to low dimensional feature space using metapath2vec. Importance of neighbouring nodes was applied using bias network projection recommendation which updates the importance score based on the degree of nodes (microbes/drugs) forming two matrices. The association scores are determined as the linear mean of the resulting matrices. [1]

By using variational graph autoencoder (VGAE), Deng et al. proposed a Graph2MDA model. A multimodal network is created consisting of a bipartite network and a heterogenous network. The node attributes were calculated based on microbe functional

and genome similarities, and drug structural and GIP similarities. Each network was fed to a VGAE - double layer GCN - to learn informative and interpretable latent representations for node attributes and whole graphs and the resulting graphs were decoded, concatenated and put through deep neural network to predict associations. [3]

Table 1 summarises the review microbe-drug association prediction techniques.

3. Data Sources

Various biological data sources were used in the construction of the heterogenous network as well as the node feature matrix.
The microbe-drug associations are obtained from the MDAD [8]. The microbe-disease associations are extracted from the HMDAD [9] for the set of microbes in the MDAD dataset. For the drugs, the drug-disease associations, are extracted from the BioSnap dataset [10], while the and the side-effect of drug combinations is taken from Bionsnap dataset [10] and study [14], for the set of drugs in the MDAD dataset. The FASTA genome sequence of microbes is downloaded from the NCBI database [15]. Further, the Stitch database [16], MeSH database [17], the Kegg databse, and DrugBank [18] database are used for mapping the drugs to IDs of each database.

4. Methodology

The workflow of the proposed methodology is presented in figure 1. It follows four main steps, first the input graph, i.e. the heterogenous network is constructed. Secondly, the node information matrix is constructuted. Next, the network and node matrix are used as input to the GNN model which encodes an embedding with prediction scores for the microbe-drug pair. Finally, the model embedding is sent as input to the DNN which predicts microbe-drug associations.

4.1. Constructing the Input Graph Network

4.1.1. Construction of Bipartite network

The bipartite network consists of undirected edges between the set of 173 kinds of microbes and the set of 1373 unique drugs from the MDAD dataset. There are 2470 known associations between the pairs retrieved from the MDAD dataset and 110 transitive associations between the pair obtained from the microbe-disease and disease-drug data sources out of which one link was overlapping with the known associations. Overall, the number of associations between the set of microbes and drugs was 2579.

4.1.2. Construction of Homogenous network

The homogenous network consists of undirected edges between the homogenous nodes. The microbe-microbe interactions are derived from study [5] which is extracted from the MIND network database [19] for 128 out of 173 microbes. The drug-drug inter-

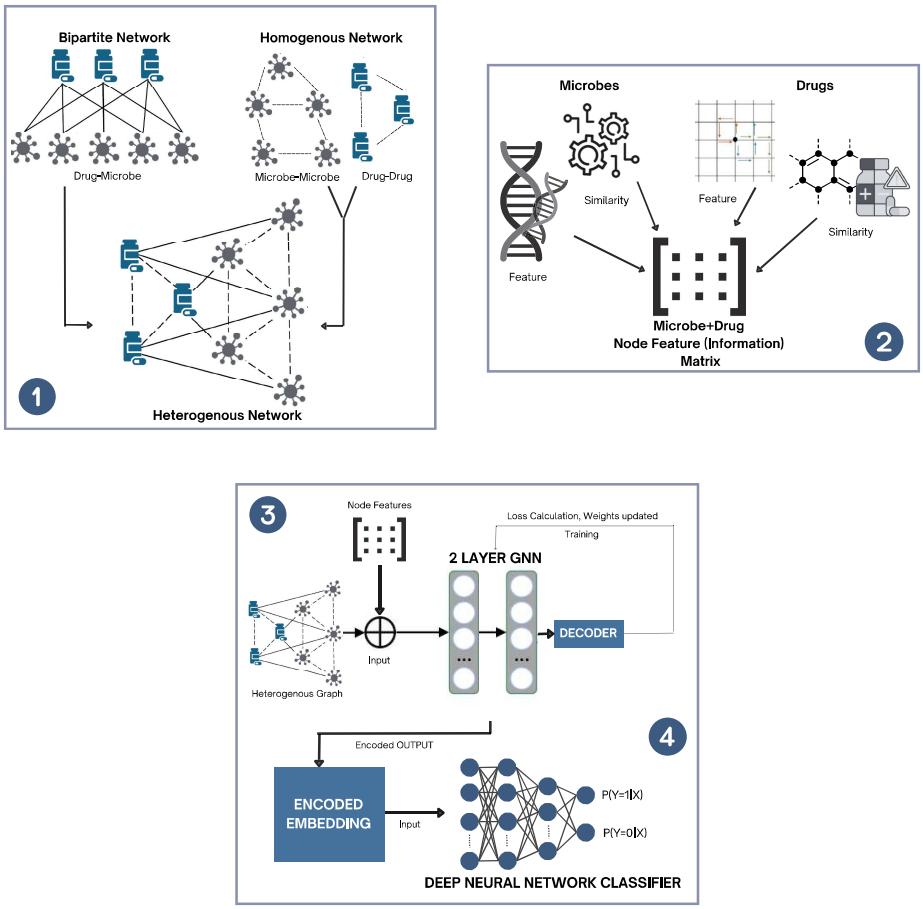

Figure 1: Flowchart of proposed methodology

actions are derived from the DrugBank database [18] for 1228 out of 1373 drugs.

4.1.3. Construction of Heterogenous network

The heterogenous network is constructed by concatenating and stacking the bipartite networks and the homogenous networks constructed. $nd + nm \times nd + nm$ is the final dimension of the heterogenous network H which is represented by equation 1 where P_m and P_d represent the homogenous network for microbes and drugs respectively and A represents the adjacency matrix i.e. the bipartite network. The constituents of H are detailed in Table 2.

$$H = \begin{bmatrix} P_m & A \\ A^T & P_d \end{bmatrix} \tag{1}$$

Network	Associations
Bipartite (known)	2470
Bipartite (transitive)	110
Homogenous (Microbes)	138
Homogenous (Drugs)	5586
Total H (w/o overlap)	10882

Table 2: Constituents of Heterogenous Network

4.2. Constructing the Node Feature Matrix

4.2.1. Microbes

Microbe-Microbe Similarity

The microbe-microbe similarity is constructed based on the functional similarities between microbes. The data is adapted from [3] and makes use of the Kamneva tool [20] for calculations. The microbe functional similarity $S_m \epsilon R^{nm \times nm}$ is calculated based on the scores connecting two microbes to by the sum of the two microbial gene families based on a protein-protein association matrix [3].

Microbe Features

The FASTA files for microbe genome sequence was extracted from NCBI [15] to be used to calculate the microbe feature matrix. The genome sequence was one hot encoded, after which Principal Component Analysis was applied to reduce the dimensionality and extract the final set of features that form the microbe feature matrix $F_m \epsilon R^{nm \times nm}$. The mean of other entries was considered for microbes that did not have any files. [3]

4.2.2. Drugs

Drug-Drug Similarity

The drug similarity matrix is constructed as a combination of Gaussian Interaction Profile (GIP) similarity, drug structure similarity, as well as side-effect based drug similarity. It is represented by $S_d \epsilon R^{nd \times nd}$.

Table 1: Summary of microbe-drug association prediction techniques

Ref	*Dataset*	*Methodology*	*Advantage*	*Disadvantage*	*Results*
[6]	MDAD, DrugBank, HMDAD, CTD, NCBI	EGATMDA	node-level and graph-level attention, efficiently preserve the importance of graph-specific neighbors and remove irrelevant noise. GCN+GAN: learn the importance of high-order neighbours in the node-level attention	Presence of noise in the features extracted	MDAD dataset, AUC: 0.959 $\pm$0.083, $AUPR$: 0.946 $\pm$ 0.011
[13]	MDAD, aBiofilm, and dataset obtained from Wang et al.'s study	GSAMDA	Takes multiple node features into account. SAE can learn unique attribute representations.	The microbe–drug association matrix is sparse and it will affect the performance of the model to some extent. Not all microbe/drugs have diseases associated with them so some defects in using microbe/drug–disease association as attribute feature exist.	MDAD dataset, AUC: 0.9496 $\pm$0.0005, $accuracy$: 0.9896$aBiofilmdataset$, $AUPR$: 0.4510 $\pm$ 0.0051
[13]	MDAD, aBioFilm, DrugVirus, and manually curated drugs and microbes from published studies.	GCN with Conditional Random Field	Leverages multiple types of prior biological information to construct similarities for microbes and drugs.	CRF models pairwise relationships and can be computationally expensive.	aBioFilm dataset, AUC: 0.952 $\pm$0.0033, $AUPR$: 0.949 $\pm$ 0.0031
[2]	MDAD	LRLSMDA			MDAD dataset, AUC: 0.9095
[1]	MDAD, aBiofilm, Drugbank, MIND	Association mining on heterogenous network embedding representation	metapath2vec possesses a powerful capability in preserving features for heterogeneous nodes and low-dimensional embedding Bias rating of nodes reflects instrinsic characteristics	Not very good at predicting associations between new drugs or microbes since it fails to attain its neighbors if there is no link in the network connecting it to other nodes. The heterogenous network is still sparse.	MDAD dataset, AUC:0.9026, AUPR: 0.91
[3]	MDAD, drugbank, MIND, NCBI	Graph2MDA	The multi-modal attribute graphs capture similarity and ontology information about drugs and microbes. GCN encoder progressively aggregates information from neighbors yielding informative representation useful for microbes (or drugs) with few annotations.	GCN doesnt take into account attention scores for a node based on neighbour importance.	MDAD dataset, AUC: 0.9732

The GIP kernel is calculated on the constructed drug-drug interaction network P_d to capture similarities between the drug making use of the formula as seen in equation 2, 3, and 4 where i and j represents drug vector at i^{th} row and drug vector at j^{th} row. Hence, the similarity matrix $D_{gip}\,\epsilon\,R^{nd\times nd}$ is constructed.

$$D_{gip}(i,j) = exp(-\gamma||P_d(i) - P_d(j)||^2) \tag{2}$$

$$\gamma = \frac{\gamma 1}{(\frac{1}{nd}\sum_{i=1}^{n_d}||P_d(i)||^2)} \tag{3}$$

$$\gamma 1 = 1 \tag{4}$$

The pairwise drug structure similarity is computed using based on the chemical structure information of using the SIMCOMP2 [21] tool. It is represented by $D_{struc}\,\epsilon\,R^{nd\times nd}$

The pairwise side-effect based similarity is calculated by mapping drugs to side effects using the drug side-effect association network from Biosnap [10], and study [14]. After mapping the similarity between a pair of drugs based on how many side-effects they have in common in $dse\,\epsilon\,R^{nd\times nd}$, extended Jaccards similarity (equation 5) was applied to calculate the similarity scores between the set of drug vectors at row i and row$j$, resulting in the matrix $D_{se}\,\epsilon\,R^{nd\times nd}$.

$$D_se(i,j) = \frac{dse(i)\dot{dse}(j)}{||dse(i)||^2 + ||dse(j)||^2 - dse(i)\dot{dse}(j)} \tag{5}$$

The fused drug similarity matrix is computed as shown in equation 6 where i and j represents drug at i^{th} row and drug at j^{th} column.

$$S_d = \begin{cases} \frac{D_{gip}(i,j)+D_{struc}(i,j)+D_{se}(i,j)}{3},\, if\; D_{struc}(i,j) \neq 0\, and\, D_{se}(i,j) \neq 0 \\ \frac{D_{gip}(i,j)+D_{struc}(i,j)}{2}, if\; if\; D_{struc}(i,j)0\, and\, D_{se}(i,j) \neq 0 \\ \frac{D_{gip}(i,j)+D_{se}(i,j)}{2}, if\; if\; D_{se}(i,j) = 0\, and\, D_{struc}(i,j) \neq 0 \\ D_{gip}(i,j), if\; D_{se}(i,j) = 0\, and\, D_{struc}(i,j) = 0 \end{cases} \tag{6}$$

Drug Features

The drug feature matrix $F_d \epsilon R^{nd\times nd}$ is constructed by applying Random Walk Restart (RWR) on the drug-drug interaction network P_d to calculate the closeness between drugs and capture topological attributes between a pair of drugs. After applying RWR, the probability distribution vector of each drug which forms the values of the feature matrix. The dataset is taken from [3] which makes use of the same set of drugs from the MDAD dataset. The formula for RWR is given by equation 7 where θ is the restart probability, and T is the transition probability matrix, $p_i^{(0)}$ is a 1D vector representing the starting probability of the i^{th} node and $p_i^{(t)}$ is a 1D vector representing the starting probability of the i^{th} node moving to other nodes at a time t. [22]

$$p_i^{t+1} = (1-\theta)p_i^t T + \theta p_i^{(0)} \tag{7}$$

4.2.3. Mutli-modal attribute construction

The feature matrices and similarity matrices constructed from drugs and microbes are stacked and concatenated with zero matrices between each of them to form $X \epsilon R^{nd+nm \times 2*(nd+nm)}$ as represented in equation 8.

$$X = \begin{bmatrix} F_d & 0 & S_d & 0 \\ 0 & F_m & 0 & S_m \end{bmatrix} \tag{8}$$

4.3. Graph Neural Networks

Graph Neural Networks are a class of neural networks that operate on graph domains. Traditional machine learning models cope with their representation by mapping it into a simpler representation. However, such preprocessing could lead to loss of important information such as topological dependency on each node, and therefore make it very hard to process such data in a simple representation. Considering that graphs are classified as complex data with underlying relationships and information represented in its structure, it must be carefully processed in order to preserve and capture structural or topological information from graphs which is where GNNs come into play. A GNN is a supervised neural network model based on information diffusion and relaxation mechanisms that directly operates on graph data. It is suitable for both graph focused, and node focused applications and opens possibility of applications in domains where the data consists of relationships and patterns. [11]

Architecture

Two GNNs namely, GCN and GAT are used to learn the latent representations from the Microbe-Drug association network and the accompanying normalized node feature matrix with information on these nodes. The architecture of the model as seen in figure 1 includes two GNN layers taking the heterogenous graph as well as the node feature matrix as inputs to encode embeddings from the information processed by the network. After the first layer, a dropout of 0.4 is applied and activation function Rectified Linear Unit (ReLU) is applied after which the output is sent to the second layer for further processing. The embeddings after the second layer are decoded by applying inner multiplication on the embeddings E to get representations R as represented in equation 9.

$$R = E.E^T \tag{9}$$

The model is updated based on the Mean Square Error Loss calculated between the matrix Y (equation 10) representing the original adjacency matrix with symmetric Laplachian normalization applied to it, and the representations R. The loss calculation is performed according to equation 11, and the model is trained on the complete dataset for 400 epochs.

$$Y = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \tag{10}$$

$$Loss_{gnn} = \frac{1}{n_d + n_m} \sum_{i=1}^{n_d+n_m} ||Y_i - R_i||^2 \tag{11}$$

The Adam optimizer with a learning rate of 0.0001 is to optimize the loss function.

4.3.1. Graph Convolutional Network

GCNs are a type of GNNs that makes use of a form of "message passing" from neighbouring nodes i.e. send input features as message from source node to its neighbours as target nodes. In this way, each node can use the information to update itself and understand the environment. The GCN consists of two layers. The first layer computes the weighted average of the messages received from the neighbourhood nodes, where the weight is based on the degree of the sender nodes, uses it as an input through a neural network model to retrieve a vector representation of the node based on the features in the environment. The output of the first layer is then used as the input to the next layer where it is processed through a neural network layer again with the output layer defined based on the task at hand. In this way, the GCN is able to capture relational structural dependencies. [23, 24] The layer-wise propagation rule of the GCN is as given in equation 12 where $\tilde{A}$ is the normalized adjacency matrix of the graph, $\tilde{D}$ is the degree matrix of the graph, $W^{(l)}$ is the weight matrix or parameter matrix at layer l, $H^{(l)}$ represents the input features at layer l,and $sigma$ is the activation function [23].

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{12}$$

4.3.2. Graph Attention Network

GATs are a type of GNNs which are similar to the GCN but make use of an extra parameter which is the attention mechanism. GCNs give the same importance to each node but GAT assigns importance to the nodes that considers the embedding of the node which can include information of local structure and features apart from just the degree of the node during aggregation. Additionally, in place of the fixed weight parameters for each node, the GAT makes use of learnable attention parameters which can be used to improve modeling capacity. In this way, the GAT is able to capture localized/context specific information in addition to relational structural information. [24, 25]

The layer-wise propagation rule of the GAT is as given in equation 13 H_i and H_j represent the feature representations of nodes i and j in the graph, W is a weight matrix, $\mathcal{N}_i$ represents the neighborhood of node i, α is a learnable attention parameter [25].

$$\begin{aligned} \alpha_{ij} &= \frac{\exp(e_{ij})}{\sum_{k\in\mathcal{N}_i}\exp(e_{ik})} \\ H_i' &= \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}\cdot WH_j\right) \end{aligned} \tag{13}$$

4.4. Deep Neural Network

The proposed methodology makes use of a simple DNN classifier as adapted from the [3] structure. The main purpose of the DNN is to take in the encoded graph embedding outputted by the GNN as input and predict association scored between the pair of microbes and drugs which will determine the presence or absence of associations between the pair.

Architecture

The DNN classifier model is a sequential dense model with an input layer, three hidden layers and an output layer of neurons. The input layer has number of neurons equal to the input target dimensions, the hidden layers have 1024, 512, and 256 neurons. After each dense layer, Batch2D normalisation, LeakyReLU activation and a dropout of 0.3 is applied. The number of nodes in the output layer is determined by the number the number of outputs or labels, i.e. number of pairs of microbes and drugs, onto which sigmoid activation is applied to identify the prediction probability of each output node which is used to classify the sample into a class. The loss is calculated between $\hat{y}$ predicted probabilities and y target variable and is used to train the model is binary cross-entropy (14) and the optimizer used is Adam with a learning rate of 0.001. The model was trained for 100 epochs with a batch size of 128.

$$Loss_{bce} = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \tag{14}$$

5. Results and Discussion

To evaluate model efficiency, statistical measures such as Accuracy, and AUC were measured. True Positive (TP) represents the number of positive samples predicted correctly, True Negative (TN) represents the number of negative samples predicted correctly, False Positive (FP) represents the number of negative samples predicted incorrectly, and False Negative (FN) represents the number of negative samples predicted incorrectly.

Accuracy (equation 15) is used to measure the total number of correct predictions out of all observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

AUC is used to quantify the capability a model has in distinguishing between classes. It calculates the area under the curve made of points formed by calculating the True Positive vs the False Positive value at different thresholds. The higher the AUC score, the better the model is at accurate prediction.

The experiments were performed using 5-fold cross validation. The results are shown in Table 3. The GCN model achieved the higher accuracy of 95.02% while the GAT model achieved the higher AUC score of 82.82%. Considering that the input to the

model was a sparse graph, the GAT was better able to predict links and distinguish between classes compared to GCN as depicted by the higher AUC score.

Model	Accuracy	Loss
GCN	**95.02**%	82.58%
GAT	94.12%	**82.82**%

Table 3: Comparative results between GCN model and GAT model

6. Conclusion and Future Scope

The proposed model was identified to be suitable to predict scores for microbe-drug associations and can extend to work for scoring known microbe - new drug and new microbe - known drug association pairs. The GCN, and GAT models produced comparable results and extensive testing is required to conclude that one is more appropriate over the other for the objective of the research.

The future scope includes adding an operation to the GNN layers to better handle the sparsity of the inputs to the GNN model and therby improve score prediction abilities. We can also further improve the model by considering the nodes of the heterogenous graph as heterogenous rather than homogenous as inputs to the GNN model. Additionally, the model architecture would benefit with the inclusion of self-supervised learning techniques due to its ability to avoid the shortcomings of overfitting and lack of generalization that comes with training a model with a large set of labeled data while providing robust prediction capabilities.

References

[1] Yahui Long and Jiawei Luo. Association mining to identify microbe drug interactions based on heterogeneous network embedding representation. *IEEE Journal of Biomedical and Health Informatics*, 25(1):266–275, Jan 2021.

[2] Lingzhi Zhu, Jun Wang, Guixiang Li, Xianglong Hu, Bufan Ge, and Bohan Zhang. Predicting microbe-drug association based on similarity and semi-supervised learning. *American Journal of Biochemistry and Biotechnology*, 17(1):50–58, Jan 2021.

[3] Lei Deng, Huang Yibiao, Xuejun Liu, and Hui Liu. Graph2mda: a multi-modal variational graph embedding model for predicting microbe–drug associations. *Bioinformatics*, 38(4):1118–1125, Nov 2021.

[4] Charisse Petersen and June L. Round. Defining dysbiosis and its influence on host immunity and disease. *Cellular Microbiology*, 16:1024–1033, 06 2014.

[5] Qing Ma, Yaqin Tan, and Lei Wang. Gacnnmda: a computational model for predicting potential human microbe-drug associations based on graph attention network and cnn-based classifier. *BMC Bioinformatics*, 24(1), Feb 2023.

[6] Yahui Long, Min Wu, Yong Liu, Chee Keong Kwoh, Jiawei Luo, and Xiaoli Li. Ensembling graph attention networks for human microbe–drug association prediction. *Bioinformatics*, 36(Supplement2), Dec 2020.

[7] Yahui Long, Min Wu, Chee Keong Kwoh, Jiawei Luo, and Xiaoli Li. Predicting human microbe–drug associations via graph convolutional network with conditional random field. *Bioinformatics*, 36(19):4918–4927, Dec 2020.

[8] Ya-Zhou Sun, De-Hong Zhang, Shu-Bin Cai, Zhong Ming, Jian-Qiang Li, and Xing Chen. Mdad: A special resource for microbe-drug associations. *Frontiers in Cellular and Infection Microbiology*, 8, Dec 2018.

[9] Hmdad online. Available online at:http://www.cuilab.cn/hmdad.

[10] Sagar Maheshwari Marinka Zitnik, Rok Sosič and Jure Leskovec. BioSNAP Datasets: Stanford biomedical network dataset collection. http://snap.stanford.edu/biodata, August 2018.

[11] F. Scarselli, M. Gori, Ah Chung Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, Jan 2009.

[12] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

[13] Yaqin Tan, Juan Zou, Linai Kuang, Xiangyi Wang, Bin Zeng, Zhen Zhang, and Lei Wang. Gsamda: a computational model for predicting potential microbe–drug associations based on graph attention network and sparse autoencoder. *BMC Bioinformatics*, 23(1), Nov 2022.

[14] Yi Zheng, Hui Peng, Shameek Ghosh, Chaowang Lan, and Jinyan Li. Inverse similarity and reliable negative samples for drug side-effect prediction. *BMC Bioinformatics*, 19(S13), Feb 2019.

[15] NCBI. National center for biotechnology information. Available online at: https://www.ncbi.nlm.nih.gov/, 2019.

[16] Available online at: http://stitch.embl.de/.

[17] Available online at: https://meshb.nlm.nih.gov/.

[18] DrugBank Online. Drugbank. Available online at: https://go.drugbank.com/, 2022.

[19] Mind-web online. Available online at: http://microbialnet.org/mind.html.

[20] olgakamneva. Kamneva tool. Available online at: https://github.com/olgakamneva/Kamneva_2016, 2022.

[21] Available online at: https://www.genome.jp/tools/gn_tools_api.html#simcomp2.

[22] Deepak Kumar Jain, Zhang Zhang, and Kaiqi Huang. Random walk-based feature learning for micro-expression recognition. *Pattern Recognition Letters*, 115:92–100, Nov 2018.

[23] Thomas Kipf and Max Welling. *SEMI-SUPERVISED CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS*. 2017.

[24] AI Overlords. Basics of graph neural networks. Available online at: https:

//www.graphneuralnets.com/p/basics-of-gnns.

[25] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *arXiv:1710.10903 [cs, stat]*, Feb 2018.